STATISTICS FOR MENTAL HEALTH CARE RESEARCH

M. Venkataswamy Reddy

NATIONAL INSTITUTE OF MENTAL HEALTH AND NEURO SCIENCES
(DEEMED UNIVERSITY)
BANGALORE, INDIA.

Community Health Cell
Library and Information Centre
367, " Srinivasa Nilaya "
Jakkasandra 1st Main,
1st Block, Koramangala,
BANGALORE - 560 034.
Phone : 5531518 / 5525372
e-mail:sochara@vsnl.com

STATISTICS
FOR MENTAL HEALTH CARE RESEARCH

Dr. M.Venkataswamy Reddy, Ph.D.
Additional Professor, Department of Biostatistics
NIMHANS, Bangalore.

NATIONAL INSTITUTE OF MENTAL HEALTH AND NEURO SCIENCES
BANGALORE, INDIA.
2002

"Statistics for Mental Health Care Research"
by Dr. M. Venkataswamy Reddy

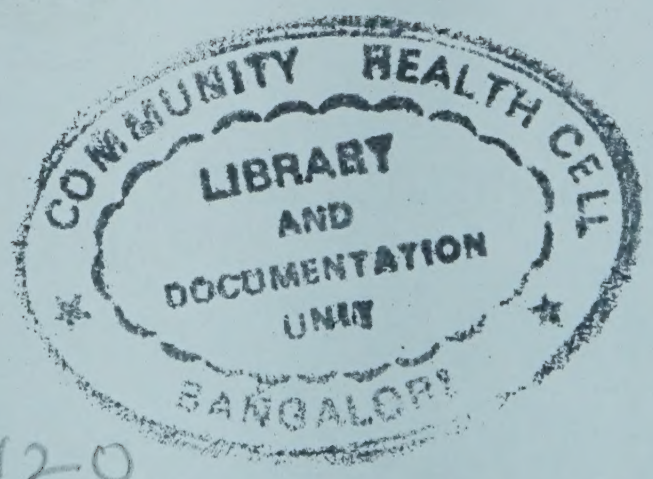
NIMHANS Publication No. : 46

First Edition : 2002

Price :Rs 200, $30.00, £ 15.00

Editor-in-Chief
Dr. M. Gourie-Devi
Director/Vice Chancellor, NIMHANS

Printed at
Anju Graphics, Tel : 6562579



Foreword

Knowledge of applied statistics is central to medical research through all the stages of planning, conducting and interpretation of data. Although there are numerous books addressing the issues in statistics for medical research, these do not cover the special aspects required by the research workers in mental health. This book "Statistics for Mental Health Care Research" by Dr. M. Venkataswamy Reddy fulfills the long felt need. It exposes the statistical methodology right from descriptive statistics to multivariate statistical methods needed for the post-graduates. All the basic aspects needed for the research worker, viz., statistical aspects of planning, conducting and analysis of a scientific investigation are also included. Although the models presented illustrate situations relevant to mental health, identical models are also applicable to communicable diseases and other medical and surgical fields. Thus, the book could have a wider medical audience too. While the author has used the matrix forms in the presentation of many multivariate techniques, the methods of approach, the solution and the interpretation has simplified these complex methods. Protocol writing, proforma for obtaining basic data of psychiatric patients, ICD-10 codes for mental and behavioural disorders, a number of tables on statistics included in the Appendix will be useful to the discerning reader.

Dr. Reddy, Additional Professor of Biostatistics of this Institution has crystallized his three decades of experience in the field and this book richly reflects his statistical and teaching skills. This book will certainly be an asset to professional and students interested in mental health research.

14 March, 2002

Dr. M. Gourie-Devi
Director - Vice Chancellor and Professor of Neurology,
National Institute of Mental Health and Neurosciences,
(Deemed University) Bangalore.

Preface

In addition to statistical thinking, the applications of statistical methods are essential in enriching one's knowledge in his subject. This is particularly true in the area of mental health care, whether it be psychiatry, clinical psychology, psychiatric social work or psychiatric nursing. Many manuals are available on these methods but none of them is precise and up-to-date for those who are in or intended to enter mental health care profession, whether it be administration, practice, education or research. This book is designed to demonstrate these methods by means of arithmetical examples which may be reworked with pencil and paper in a matter of minutes. Most frequently the simplicity of the methods are maintained by a suitable choice of artificial data. High-speed electronic computers are available to aid in the analysis of substantial amount of actual data. The real data for illustrations are taken from my own work and publications. It is easy for the reader to understand important formulas by explaining themselves on common-sense grounds, the roles played by different parts of the formula. The investigator has to find out which statistical method is appropriate for his investigation goal, and in my opinion the converse is also true 'the investigator must be familiar to a wide variety of basic statistical methods in order to state his problem clearly and to generate hypotheses'. This book is intended to be self-contained and covers the major topics and applications of statistics. The major features of this manual are the development of mental hospital service indicators in chapter 4, and inclusion of the ICD-10 diagnostic categories of mental and behavioural disorders as an appendix. The neurologists and basic medical scientists will have no difficulty in translating the data of patients into terms which are more familiar within their own disciplines.

NIMHANS, Bangalore
February, 2002

Dr.M.Venkataswamy Reddy

Contents

CHAPTER 1

INTRODUCTION

There are two meanings that can be attached to the term *statistics.* The first refers to the facts and figures of any kind. Thus, we speak of population statistics, mental health statistics, medical statistics and in general bio-statistics. It also refers to a body of knowledge known as statistical methods which are based on the theory of probability. These are the mathematical devices of use in discovering certain differences, trends, relationships, probable predictions and distinct groups from the mass of data. They aid the experimenter in uncovering facts that may not be immediately obvious in his mass of data. They help the researcher in deciding whether differences between groups, trends of factors, relationships between variables, accuracy of predictions, and distinguishability of groups are large enough to be considered significant. They serve their purposes when used in conjunction with experimental studies, observational studies or experimental enquiries. The scope of statistics may be briefly specified as: (1) organization and collection, classification and presentation, and summarization of data, (2) estimation of parameters and tests of significance, (3) formation of distinct groups and identification of individuals, and (4) presentation of complex multivariate data in a simplest form.

Health is a state of complete physical, mental, social, and spiritual well-being and not merely absence of disease or infirmity. *Psychiatry* is a branch of medicine which deals with the recognition, treatment and prevention of mental disorders and abnormalities. It is generally felt that the human mind is greatly influenced by the environment, social and family surroundings, education, cultural milieu and the biochemistry of the individual. As in the case of general medicine, the field of psychiatry may be broadly classified into clinical psychiatry and community psychiatry. Statistics have vital role to play in both clinical settings and community settings. In *clinical psychiatry,* statistics are used in the areas of: (1) documentation of natural and medical history of various psychiatric disorders, (2) definitions of abnormal conditions, (3) planning of clinical trials, (4) evaluation of different levels of treatments, (5) providing standard measures of accuracy of various clinical procedures, and (6) predicting the outcome of common psychiatric disorders. In *community psychiatry,* statistics are widely used in the areas of: (1) assessment of the state of mental health in the community, (2) indication of the basic factors underlying this state of mental health in the community, (3) planning and monitoring of mental health programs for specific population groups, (4) evaluation of the total program of action, and (5) promotion of mental health legislations.

(a) The Scientific Approach

Statistics are the tools of science to deal with the mass of data. The entry of psychiatry into the scientific world is recent. In order to be called as a scientific discipline, the field of psychiatry has also undergone the rigors of *various approaches* to acquire knowledge such as the method of tenacity (blind belief), the method of authority (established belief) and the method of intuition (a priori method). In method of intuition, obvious truth is claimed for propositions. Here, a priori propositions are agreed by reasoning and not necessarily with experience

such as females are more prone to most of the mental disorders, urbanization is not good for mental health etc. It is quite possible for two persons to arrive at different conclusions using these approaches. Hence, we *need a method* which when two mental health workers use it, both must arrive at the same conclusion with the same data in making a diagnosis, deciding on the course and method of treatment, predicting the outcome of treatment, and in determining the cause of a mental disorder in order to prevent it. That method must include the contradictory evidences, free from human bias, and must be based on empirical testing. Such is the method of science. The scientific method is the application of logic and objectivity to understand the phenomena. The science is the philosophy to search for knowledge (research) in mental health care. The methods and techniques used in research methodology are given the name of the scientific method.

There are *five steps* involved in any scientific study, viz., the formulation of problem, formulation of hypotheses, modification of the problem/ hypotheses, testing (experimentation or verification) and generalization. During the contact with the field of study, the scientist may experience an obstacle to understand a phenomenon. Hence, his first step is to express the problem in some manageable form. A *problem* is a question proposed for solution. Usually, a problem arises in two occasions. Firstly, to develop a theory to explain an event which is occurred and none of the existing theories explained it so far. Secondly, to modify the existing theory so as to suit the new event. True scientific enquiry has no beginning and has no end. The fact gathering activity is an intermediate stage in scientific enquiry. The scientist must have good acquaintance with the area in which he has to identify the problem. The subject matter for mental health care research is different from that of the biological or physical sciences: the manifestations of psychiatric disorders cannot easily be quantified, it is difficult to measure the degree of severity of psychological disturbance and to determine

its time of onset and duration. The correct statement of the problem sets the direction of the study, reveals procedures and methods, and aids in control of bias.

The formulation of *hypotheses* means to give tentative answers to the problem under study. This can be done only after turning back on the experiences for possible solutions to the problems and after examining different phenomena. In the given circumstances and situations, the scientist may alter or *modify* the problem or the hypotheses that he has formulated. Here he may use the deductive reasoning. After these three steps, the scientist *tests* the validity of the various answers that he has suggested to the problem empirically. Finally, on the basis of having established that certain factors are responsible for the occurrence of the phenomenon, the scientist makes some *general inferences,* conclusions, principles, theories or rules. Only from these propositions, certain statements relating to the specific occurrences of the phenomena are deduced. In mental health care research, hypothesis about the etiology of mental illness are so diverse as to defy incorporation into any comprehensive theory

The scientific method has several *specific features* such as the circularity aspect of *facts and theory,* humility, free enquiry, free from human bias, controlled observation, comparability, repeatability, ruling out metaphysical explanations, and use of both inductive and deductive reasoning. A fact is an empirically verifiable observation, whereas theory is the relationship between facts. Theory is the tool of science because it predicts facts and points the gaps in our knowledge. On the other hand, facts are also productive of theory because facts help to initiate, reformulate, clarify and change the focus and orientation of theory. Thus, facts suggest theory, the theory suggests that certain other facts be made to test it, the new facts modify the theory, and the modified theory suggests still other observations. For example, the identification and treatment of psychotic depression has undergone several changes

due to the modifications of the syndrome and discovery of new antidepressant drugs. The scientist takes into account of *every piece of knowledge* available to him. The scientist is *entirely willing* to have his hypothesis disproved by others. In fact, he tries as hard as he can to have his hypothesis disproved. The theory must be established by methods free from *personal bias.* Our own bias and interest may enter even in the choice of research that we are undertaking. We choose it either because we have a special interest in it or we want to avoid some other aspect that we cannot face for some reason.

The *controlled observation* is essential for drawing valid conclusions. When a hypothesis seems to be supported by an experiment, the scientist will test the alternate hypothesis also. If the alternate hypothesis will also be supported, then he will doubt the hypothesis that he has suggested for the study. In other words, if a scientist finds that A goes with B, then he must also test whether A goes with non-B. For instance, one should not make the diagnosis of schizophrenia based on the presence of hallucinations only because hallucinations may also present in psychotic depression. Since the scientist is always willing to have his hypothesis disproved by others, the data should be collected in ways *comparable* with that of others. This requires that a detailed description of the study is essential and when reported, there will be no secrecy in the material. The diagnostic criteria for paranoid schizophrenia given in the tenth revision of the International Classification of Diseases (ICD-10) of the World Health Organization, differs from the one given in the fourth edition of the Diagnostic and Statistical Manual (DSM-IV) of the American Psychiatric Association. A piece of research work undertaken by a scientist should be *replicated* by others. Only then, one can become sure that the relationships established are general and not unique.

The scientists are *avoiding meta-physical explanations* while carefully attempting to explain the relationship between various facts.

A meta-physical examination is simply a proposition that cannot be tested empirically such as the Id, Ego and Superego theory of Sigmund Freud. What were meta-physical issues a few decades ago have now become accepted sciences such as placebo effect, expectancy effect and mind-machine interaction. Moreover, a true scientist can never offered to say 'this is not true'. A true statement may be made like 'I do not have means to study this' or 'I am not convinced of the relationship between data and inferences about this' etc. The scientist looks for reasons while making an inductive or a deductive inference. An *inductive inference* is concerned with drawing valid conclusions on the population based on the information contained in the sample. A new anti-psychotic drug administered to a sample of schizophrenic patients was found to be effective when compared with the placebo group. Then it is possible to conclude that the drug is effective to schizophrenics. A *deductive inference* is the science of drawing valid conclusions on a particular sample based on the information contained in a reference or a target population. It is established that the prevalence of mental and behavioural disorders in India is 58.2 per one thousand general population. It seems reasonable to conclude that the prevalence rate for Kerala state is the same as that of India. It does not seem reasonable to conclude that the general structure of Indian families is the same as that of the Kerala families.

(b) Mill's Canons

Statistics are used to *determine causal factors* of phenomena in experimental enquiries. Sir Francis Bacon was one of the first men who wrote about logical devices that could be used to aid the researcher in his search for scientific knowledge. John Stuart Mill undertook the task of polishing the general concepts of Bacon and presented the logical devices as short rules under the title 'methods of experimental enquiry'. These rules indicate five canons, viz., the method of

difference, method of agreement, joint method of difference and agreement, method of concomitant variation and the method of residues.

The *method of difference* consists in making use of two groups of subjects equal in all respects and do something to one of the groups (experimental group) and not doing any thing to the other group (control group). If a change takes place in some dependent variable in the experimental group, but does not take place in the control group, then the change in the dependent variable is attributed to the manipulation in the experimental group. That is, if

JK → LM (control group)
JKC → LME (experimental group)

then C is related to the occurrence of E. For example, sixty schizophrenic patients were randomly divided into two groups of 30 patients in each group. One group was given educational input along with the treatment and the other group was given the treatment without additional educational input. The educational input group has demonstrated better understanding of the illness which led to better follow-ups and management. Such experiments are frequently carried out in mental health care research field. Thus it is not a method of discovery since the researcher chooses his independent variable because he believes that it may be related to the dependent variable and not for any other reason. This is not a method of proof since the cause of an event may be due to a multiple factor. It is difficult to conclude that two groups are equal in mental health care research. The method allows the experimenter to verify his hypothesis concerning the importance of his independent variable in the occurrence of the phenomenon. There is no better method of experimental enquiry when this method is applicable.

The *method of agreement* consists in observing the occurrence of the phenomenon in each time where a specific independent variable presents, a specific dependent variable occurs. The researcher continues to observe the occurrence of the phenomenon with different combinations of his independent variable present until he ascertains that there was only one specific independent variable always presents whenever a specific dependent variable occurred. Then he would infer that the independent variable that was common to the occurrence is related to the occurrence of the dependent variable. That is, if

ABC → FGE
BDC → GHE
ADC → FHE

then C is related to E. For example, the case history records of 28 conduct disorder children revealed that all of them had either inadequate parental control or family over involvement. By applying the method of agreement, it is possible to state a relationship between the occurrence of conduct disorder and the presence of inadequate parental control or family over involvement. Thus this method yields evidence that was previously suspected and hence it is not a method of discovery. In a more complicated situation wherein we have little knowledge of all the factors present, the wrong conclusion may be reached through the use of this method. However, it helps in formulating hypothesis regarding the cause of a given event and provides a means of verifying it by a more systematic methodological approach.

The *joint method of difference and agreement* consists in testing to see that if in two or more instances of occurrence of the phenomenon have only one factor in common and then test to see that if in two or more instances where that common factor is absent, the phenomenon does not occur. Then the researcher concludes that the common factor is related to the occurrence of the phenomenon. That

is, if in

First instances	:	ABC	→	JKE
		DFC	→	LME
		GHC	→	NOE
Second instances:		PQ	→	VW
		RS	→	XY
		TU	→	Za

then C is related to E. For example, the case history records of 200 child guidance clinic children revealed that all the 23 children with hyperkinetic disorder had abnormal attention and concentration and only two of the remaining children had this abnormality. Hence, it is possible to say that certain relationship may exist between the occurrence of hyperkinetic disorders and the presence of abnormal attention/concentration in children. But it is doubtful to say whether there is only one factor in common in the occurrences of the phenomenon. It overcomes some of the difficulties faced in the above cited methods.

The *method of concomitant variation* consists in recording the variation of both the independent variable and the dependent variable, and if the dependent variable varies in any manner whenever the independent variable varies in some particular manner, then the experimenter concludes that the two variables are related. That is, if

AB (1C) → DF (1E)

AB (2C) → DF (2E) etc

then C is related to E, or is connected with it through some fact of causation. The findings of mental morbidity studies carried out in India revealed that several relationships between the prevalence of mental/behavioural disorders and several biosocial variables. They indicate that the risk of mental and behavioral disorders of a person

decreases with increase in socio-economic status, the risk increases gradually up to 40 years of age and decreases thereafter, the risk increases with both the size of his family and urbanization of his locality etc. Thus this method does not indicate which is the cause and which is the effect any more than the rest of the canons. The method establishes relationships between variables, suggesting cause and effect relationship. It is a method of proof when stated negatively 'nothing can be the cause of a phenomenon which does not qualitatively vary where the phenomenon varies'.

The *method of residues* consists in attempting to determine, through experimentation and deduction, that specific and identifiable dependent variables occur in a phenomenon are due to the effect of the presence of specific and identifiable independent variables. The researcher continues to ascertain such information until the relationship between one dependent variable and one independent variable is unknown in the situation. Then the investigator infers that this remaining independent variable is related to the remaining dependent variable. That is, if it is known that A causes B, D causes F, and G causes H in the paradigm ADGC → BFHE and that there are only C and E remains, then the following statements can be made:

A	→	B	
D	→	F	
G	→	H	known
C	→	E	Inferred

For example, a series of five drugs are administered to a group of patients and the responses are noted down. Then each drug is withdrawn every time from least importance to most importance in order to see at which withdrawal of the drug the patients stop improvement or regressing. Thus the method requires the experimenter

as to know where to look for the cause of the phenomena. If the cause of a phenomenon is a complex factor whose force as a causative agent disappears when one after another of its elements is eliminated, then this method does not apply. The method allows one to approximate the area within which the cause of the phenomenon rests.

(c) Types of Study

The problems in mental health care may be investigated in a variety of different ways. The approach depends on the *type of study.* The studies may be broadly classified into experimental and observational. An experiment is a study in which the investigator deliberately sets one or more factors to a specific level. A laboratory experiment takes place in laboratory where experimental manipulation is facilitated. A comparative experiment compares two or more treatments. An experiment is a cross-over experiment if the same experimental unit receives more than one treatment. The different treatments are given during non-overlapping time periods. A clinical study takes place in the setting of clinical medicine.

An *observational study* collects data from an existing situation. The data collected does not intentionally interfere with the running of the system The observational studies may be cross-sectional or longitudinal. A cross-sectional study collects data on study units at some fixed time. A longitudinal study collects information on study units over a specified period of time. The longitudinal studies may be retrospective or prospective. A retrospective study is one in which individuals having a particular outcome are identified. A case control study selects all cases usually of a disease, meeting fixed criteria. A comparison group for the cases, called controls is also selected. The cases and controls are compared with respect to various characteristics. In a matched case control study controls are selected to match characteristics of individual cases. The cases and the controls are associated with each other.

A prospective study is one in which a cohort of people is followed for the occurrence of specified events. A cohort is a component of a population identified so that its characteristics can be ascertained as it ages through time. The ideas of these studies are expanded in the following chapters.

CHAPTER 2

ORGANIZATION AND COLLECTION OF DATA

The collection of relevant and reliable data is the first and foremost important step in any scientific study. No methods of statistical analysis can compromise with the badly collected data. Prior to the collection of data, it is convenient to *write a protocol* specifying the complete plan of study and analysis. The contents for writing a protocol are indicated in Appendix I. These steps are expanded in latter portions of the book.

The study starts with the formulation of problem and review of literature. Need for the study is an important consideration as it specifies the priority to be given and the amount of money required for the study. The survey may be for operational purpose such as planning of mental health services and evaluation of treatment, or for a more academic nature such as investigation of factors affecting the origin and course of mental illness. It may be for determination of mental and behavioural disorders in the community, rising trend of substance use disorders, optimum bed strength for a mental hospital and so on.

In case the sanctioned budget is sufficiently large, the researcher collects the primary data for the purpose of the study. The secondary data is the data extracted from records maintained by other agencies for some

other purposes. The major sources of secondary data for mental health care research are psychiatric records of government mental hospitals, general hospital psychiatric units, private psychiatry nursing homes and clinics, and specific surveys on mental morbidity etc. The records of treatments are rarely maintained by traditional healers such as Mantrawadies, Ayurvedic healers and temple priests in India.

The investigator has to specify clearly the aims and objectives of the study. The *objectives* of the study are to discover answers to questions through the applications of the scientific procedures. According to the nature of the objectives, the studies may be classified as *exploratory* to gain familiarity with a phenomenon, *descriptive* to portray accurately the characteristics of a particular group, *diagnostic* to determine the frequency with which something is associated with something else or *hypothesis-testing* to test a hypothesis of a causal relationship between variables. The *aim* of the study is to find out the truth which is hidden and which has not yet been discovered.

(a) Population, Sample, Sampling Units

The term *population* has wider meaning in any research activity. It is the aggregate of units of observation about which certain information is required. On the other hand, a *sample* is a part or a portion of the population selected in scientific manner. When the investigation is carried out for all the units in the population, it is called as census enumeration. When the investigation is carried out for a sample of units, it is called as a sample enumeration. The statistical constants based on population values are known as parameters and those based on sample values are the estimates of the parameters. The population may be finite, infinite or hypothetical.

The individual subjects upon whom the data are collected are known as *sampling units.* In an investigation, the sampling unit has to be clearly defined. The sampling unit may be an institution, medical

college, mental hospital, psychiatry nursing home, psychiatric clinic, village, family or an individual. If the sampling unit is an individual, then the definition should give clearly as to what type of individual need to be included. In mental morbidity surveys, the basic data of all the individuals and again psychiatric data of identified cases in selected families have to be collected. The general principle governing the definition of the sampling unit is that all the units taken together constitute the population and there should be no overlapping of units.

(b) Variables, Levels of Measurement, Standardization of Terms

The characteristics on which the observations are made are known as *variables.* The variables may be qualitative or quantitative. In qualitative data, there is no notion of magnitude or size of the attribute as the same cannot be measured. A *qualitative* variable may be dichotomous (for example, gender, and signs such as stupor and tremors) or polychotomous (religion). The quantitative data have a magnitude. A *quantitative* variable may be discrete (a family size) or continuous (age). The choice of biosocial variables to be included in a study is an important consideration and this may be done by consulting an expert in the field.

The researcher has to measure all the variables included in the study. *Measurement of variables* is the assignment of names or numerals to these variables or observations. In a physical science measurement, all the basic arithmetic operations, viz., the addition, subtraction, multiplication and division may be made use of. But the measurement in behavioural sciences are complex and all these operations cannot be made use of. Basically, four levels of measurement have been established, viz., the nominal, ordinal, interval and ratio levels of measurement. Measurement in which a name is assigned to each observation belongs to the *nominal scale* of measurement. The variables such as gender (the categories are male and female), religion (Hindu, Muslim,

Christian, Sikh, Jain, Buddhist, etc.) and diagnostic groups (organic psychoses, substance use disorders, schizophrenia, affective disorders, neuroses, behavioural syndromes, personality disorders, mental retardation, developmental disorders, and behavioural/ emotional disorders of children) are in nominal scale. This scale simply involves in the classification of subjects according to specified categories. The nominal data is also called as categorical data or enumeration data. The patient registration numbers, coding key numbers etc., are only conventional symbols for names. The categorical data analysis involves in numbers representing frequencies. This is the crudest level of measurement.

The *ordinal scale* differs from the nominal scale in that it ranks the different categories specified in the scale in terms of a graded order. The variables such as socio-economic status SES (specified as high, middle, low), mental retardation (borderline, mild, moderate, severe, profound) and prognosis (recovered, improved, slightly improved, not improved) are measured in ordinal level. The high SES is better than the middle SES. Obviously, the high SES is better than the low SES, but we do not know how much better. That is, the distances between the successive scale points may not be equal in ordinal scale. The mental retardation may be classified according to intelligence quotient as borderline (70-90 IQ), mild (50-69), moderate (35-49), severe (20-34) and profound (below 20 IQ). The mental health care research is mainly concerned with nominal data, ordinal data and ranking data which are generally called as *qualitative data.*

An *interval scale* has all the characteristics of an ordinal scale and in addition, the distances between the successive scale points are of equal size. Here, the zero point and the unit of measurement are arbitrary. The variables such as temperature measurement with centigrade thermometer, IQ and achievement test scores are in interval scales. The 0^0C is not the absence of temperature; it is the temperature at

which water freezes into ice. A *ratio scale* has all the characteristics of an interval scale and in addition, has an absolute zero point, a point at which the variable being measured is totally absent. Here, the ratio of any two scale points is independent of the unit of measurement. The physical variables such as height and weight are measured in ratio scale. All the four basic arithmetic operations are permissible in ratio scale, while only addition and subtraction are permissible in an interval scale. The quantitative variable may be measured either on an interval or on a ratio scale. On the assumption that no serious errors will be incurred, the scores obtained on psychosocial variables may be treated as if they are measured in ratio scales and subjected to suitable statistical methods. But such a generalization on an ordinal scale is a serious error. The data in interval scale and ratio scale are generally termed as *quantitative data.* The data measured in ratio scale may be converted into interval data, the interval data can be converted into ordinal data, and the ordinal data can be converted into nominal data.

All the variables, categories and terms used in the study must be clearly defined and *standardized.* The Indian Council of Medical Research has defined an urban locality as one with closed drainage facility, semi-urban as one with open drainage facility and rural as one with no drainage facility. In mental health information system, it is convenient to consider the deaths, suicides, homicides, against medical advice and escapes as specific categories of discharge.

(c) Proforma

The investigator collects the data by using a proforma which may be a questionnaire, schedule or record forms. In the *questionnaire* approach, the respondents fill the proforma and return them or send them by post. In the *schedule* approach, the investigators ask specific questions and fill the proforma. The *record forms* are used to extract secondary data. The questionnaire approach may result in high non-response.

The schedule approach is costly, time consuming and labour intensive, and may carry considerable systematic response errors. Some times, the data may be obtained by observation. The telephone interviews may be preferred in case the survey has to be accomplished in a very limited time. The list of demographic and socio-economic variables are to be kept in the first section of the proforma followed by the tools to be used and the list of clinical investigations to be made in the subsequent sections. Prior to the finalization of proforma, it is advantageous to carry out a *pilot study* on a small number of units. It is convenient to make the proforma pre-coded especially in case of large scale surveys. A well tested and pre-coded proforma to collect basic data of psychiatric patient who has been treated in a mental hospital is given in Appendix II.

The *three-digit* ICD-10 *codes* for mental and behavioural disorders categories are given in Appendix III. The codes for occupational categories are given below:

00 Unemployed
01 Professional
02 Administrative/Executive/Managerial etc
03 Clerical
04 Business
05 Farmers/Coolies/Fisheries etc
06 Mining/Quarrying
07 Transport/Telecommunications
08 Industry/Productive Workers
09 Service/Recreational/Sports etc
10 Aged/Retired
11 Students
12 Housewives
13 Non-agricultural Labourers
14 Household Girls
15 Age below 7 years.

(d) Construction of Rating Scales

The measurement of psychosocial variables may be accomplished by using the methods of rating scales. Scaling is the process of developing standard measurements whereby individuals may be compared with

regard to their attitudes. The simplest way of measuring one's attitude is to ask him to rate his strength by himself. This can be done by presenting him with a number of attitude statements of varying intensity (Likert scaling) or an hypothetical range of attitudes from extreme favourableness to extreme unfavourableness graphically or pictorially (Thurstone scaling).

In **Likert scaling method**, the subjects are provided with a list of items (statements) with which they may agree or disagree to varying degree of intensity. The typical Likert scale items and response categories are illustrated below.

Attitudes of Superintendents towards an Ideal Medical Records Section for a Mental Hospital

Items	Response categories				
	Strongly agree	Agree	Don't know	Disagree	Strongly disagree
1. It is necessary to maintain individual case files for psychiatric patients	–	–	=	–	–
2. Maintenance of case records for more than ten years is a burden to the hospital	–	–	=	–	–
3. The number of missing case records should not exceed two percent	–	–	=	–	–
4. The time taken to retrieve a case record should not exceed ten minutes	–	–	=	–	–
5.					

The investigator has to take a decision as to how many items have to be selected from the bank of items. About fifty items have to be selected after eliminating those of duplicity and ambiguity. These items must be valid, reliable, discriminative and moderately difficult. The items are listed in such a way that the responses have to be spread widely. Usually, 3 to 5 response categories are to be employed. If it is too large, then the respondents will find difficult to place themselves

on the scale. With an odd number of response categories, there will be a middle category representing the neutral position.

The intensity of a given attitude is determined in part by weighing the response categories to each item in the scale. Weighing is the process of assigning numerical values to each of the response categories. The response categories for the items measuring attitudes towards an ideal medical records section may be given the weights of 5, 4, 3, 2, 1 or 1, 2, 3, 4, 5 respectively. Logically, larger the score more the respondent has positive attitude. The respondent attitude is determined by summing the scores for all the items in the scale, and hence it is also called as summated rating scale. Some of the items on the scale are negatively worded and the remaining are positively worded. In order to make the total score meaningful, positive items must be scored in one order and the negative ones in the reverse order. It is advantageous to have roughly equal number of positive and negative scored items in the scale.

The summated rating scale is the commonly used scaling method since it is easy to construct, to score and to interpret. Summated rating measures lend themselves to ordinal level of measurement. There is no consistent meaning that can be attached to the raw scores derived by such measurement. It is assumed that each item has the same weight in comparison with those of other items and this may not be a valid assumption. Persons receiving the same score on this measure may not possess the traits to the same degree.

The **Thurstone scaling method** consists in assigning a specific scale value to each item in the scale that stands for the intensity with which it measures what it is supposed to measure. Consider the following two items measuring the attitude towards Christianity.

A: I would like to have a Christian as my neighbour

B: I would like to marry a Christian

A person who agrees with the item A may not agree with the item B. But a person who agrees with item B will almost certainly agrees with the item A. The two items measure the same attitude, but they differ in terms of their intensities. The table of affixing numerical value to each item is accomplished through the use of judges. About fifty items are selected from the bank of items. About thirty judges are asked to sort out these items into about seven categories ranging from high to low intensity. Judges whose sorting indicates that they have failed to do their job perhaps due to misunderstanding of instructions or just carelessness are to be eliminated. The weight of each item is based on the average of the categories into which it has been sorted out by the judges. After these procedures, about twenty items are selected for use in the final questionnaire by cutting out those with high scatter. The scale must cover all the ranges of the attitude. Now the respondents are asked to select two or three items that best reflect their sentiments towards the psychological object in question. We calculate either the mean or the median of these items to portray their positions along the continum of intensity. The Guttman scalogram analysis may be used to determine whether all the items measure the same dimension, that is whether the scale is unidimensional or not.

A large number of individuals may be compared with regard to their attitudes with Thurstone scaling. The judges eliminate bad items or the items with no consistency. The Thurstone scaling measures lend themselves to interval level of measurement. There is no way of controlling the influence of the judges bias in sorting. Again, the attitudes of the judges who sort out the items may be quite different from those of the respondents whose attitudes are to be scaled. It is possible to get identical scores with this scaling method.

(e) Response Errors, Non-response Errors

The *response errors* are the wrong or biased answers. They are mainly due to self interest or false prestige of respondents, and stigma attached

to mental disorders. Some times, these errors are accidental. The response-errors may be reduced by consistent definitions, tools and measurement process. These errors sometimes called as outliers may be easily identified by preparing a master chart and screening the data by internal consisting check, as shown in the following master chart. For example, the age of 64 years is not consistent with the diagnosis of dementia in Alzeimer's disease with late onset of the first patient. In the master chart, the rows represent the sampling units and columns represent the variables. This refers to the editing of data which is a process of examining the collected raw data to detect errors and correct these when possible. This is to assure that the data are accurate, consistent with other facts gathered, uniformly entered, as complete as possible and have been well arranged to facilitate coding and classification.

Patient name (Initials only)	Registration number	Age (years)	Gender	Diagnosis (ICD-10)
1) Mr.SGR	6793	64	male	F00.1
2) Mr.DB	6794	37	male	F20.0
3) Mr.MVJ	6795	38	female	F20.3
4) Mr.BPS	6797	28	male	F30.1
5) Mr.KDR	6798	31	male	F53.0

The *non-response* means failure to measure some of the units in the selected sample. Some individuals may be temporarily away from the house at the time of the survey, they may not have information wanted in certain items in the proforma or they may not like to give their personal informations. The non-responses include non-coverage also. The non-coverage may be due to incomplete sampling frame, insufficient budget, poor transportation facilities or unable to meet the units during the period of collection of data. They may be

minimized through adequate planning, training, monitoring and supervision. The non-response group has to be compared with the response group and the discriminative variables may have to be identified. Some times, the bias due to non- response remains indeterminate. A standard procedure is to carry out a call back on a sub-sample of non-respondents and the information thus obtained may be used to estimate on the non-response group. The proportional allocation model may be used in this procedures.

CHAPTER 3

CLASSIFICATION, PRESENTATION AND SUMMARIZATION OF DATA

In case the data is collected for a few number of units, it may be presented as it is in the text of the report. In large data sets, the observations on each variable have to be classified and presented in the form of tables. The chief objectives of classification are to simplify the complex nature of the data, tailoring it up and make it easy to grasp. It helps in making comparisons and drawing conclusions without considering directly hundreds of individuals. The classification by person may be made according to qualitative or quantitative aspects. In a *qualitative classification,* a natural classification system may be available. In classification of patients according to gender, it is enough if we count the number of males and the number of females. In a *quantitative classification,* the optimum number and lengths of frequency classes depend on the size and nature of the data, and the purpose of the study. Generally, the number of frequency classes will be between three and ten. If it is too large, then the purpose of classification many not be served. If it is too small, then some salient features of the data may be lost. Generally, the frequency classes are of equal length, mutually exclusive and exhaustive. In order to compare with the census figures of the general population, it is convenient to classify psychiatric patients according to the age groups of 0-9, 10-19, ..,70-79 and above 79 years.

Table 3.1 Demographic and Clinical Characteristics of Psychiatric Patients of NIMHANS Hospital

Characteristics		Number (500)	Percentage (100.0)
1. Age (completed years):			
	0-9	59	11.8
	10-19	67	13.4
	20-29	125	25.0
	30-39	121	24.2
	40-49	84	16.8
	50-59	28	5.6
	60-69	12	2.4
	70-79	3	0.6
	Above 79	1	0.2
2. Gender:			
	Males	303	60.6
	Females	197	39.4
3. Duration of Illness:			
	Below 1 month	58	11.6
	1 - 6 months	119	23.8
	0.5 - 1 year	56	11.2
	1 - 2 years	92	18.4
	2 - 4 years	91	18.2
	4 - 20 years	84	16.8
4. Diagnostic Blocks (ICD-10):			
	Organic Psychoses (OP)	14	2.8
	Substance use Disorders (SUD)	58	11.6
	Schizophrenia (Sch)	124	24.8
	Affective Disorders (AD)	139	27.8
	Neurotic Disorders (ND)	74	14.8
	Behavioural Syndromes (BS)	5	1.0
	Personality Disorders (PD)	3	0.6
	Mental Retardation (MR)	58	11.6
	Developmental Disorders (DD)	10	2.0
	Behr./Emotl.Disorders (BED)	15	3.0

Source : The author's study on biosocial characteristics of psychiatric patients of NIMHANS hospital during the year 1999.

(a) Tabulation

The *classified data* has to be presented in the form of tables separately for each variable or for groups of variables. The total of the frequencies, the relative frequencies, and the cumulative frequencies in the table will aid in the interpretation of the results. The table must be self explanatory. Appropriate title must be given at the top of the table. The source of data and new terms used must be specified at the bottom of the table. The classification according to four variables of a sample of psychiatric patients consecutively registered at NIMHANS hospital, are presented in Table 3.1.

(b) Graphs, Diagrams, Spot Maps

In more complex situation of presentation, further to see the shape and pattern of frequency distributions, graphs and diagrams may be used with or without the frequency tables. The graphs and diagrams prove nothing but bringing out the salient features readily in the data. They are good visual and mental aids for both professionals and laymen. In order that they present ideas truthfully and emphasize correct ideas, they must be drawn following certain basic rules which one depends partly on convention, partly on mathematical conditions and partly on personal grounds. Generally, the diagrams are drawn for qualitatively classified data and the graphs are drawn for quantitatively classified data. In *pie diagram,* the circle is divided into segments with their angles proportional to the frequencies. Here, the 360^0 are equivalent to the total of the frequencies. This diagram is more appropriate for nominal data as shown in Figure 3.1. The load of the male psychiatric patients in comparison with that of the female patients at NIMHANS hospital may be clearly viewed from the diagram.

In *bar diagram*, on a common base line, vertical or horizontal bars are drawn with their heights proportional to the frequencies.

The bars are of equal width and successive bars are drawn at equal distances. Composite, component or percentage bar diagrams may be drawn depending on the size and nature of the classified data. This diagram is especially preferred when time is one of the components. It is appropriate for discrete data as shown in Figure 3.2. The extent of under representation of five diagnostic groups (SUD, ND, PD, MR, BED) at NIMHANS hospital service, may be clearly viewed from the bar diagram.

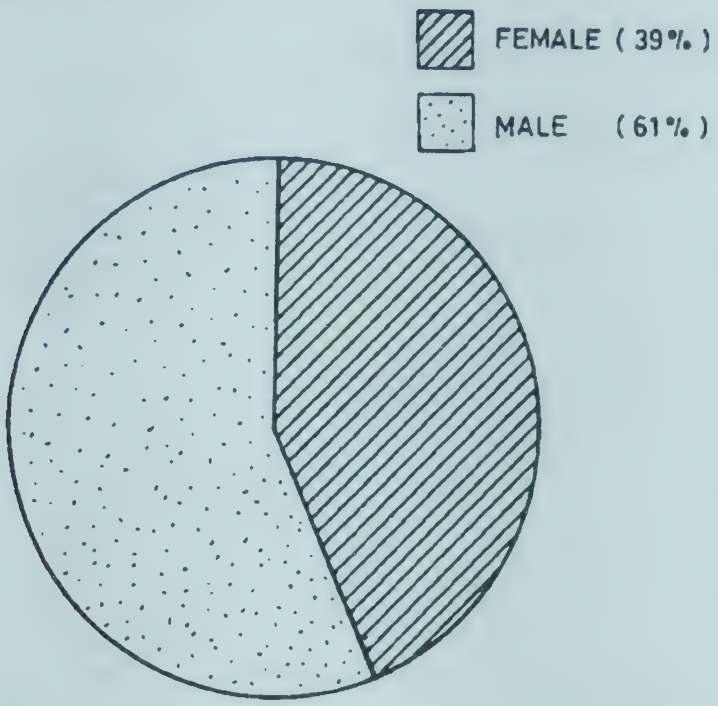


Figure 3.1 Pie Diagram Showing Gender Distribution of 500 Psychiatric Patients of NIMHANS Hospital.

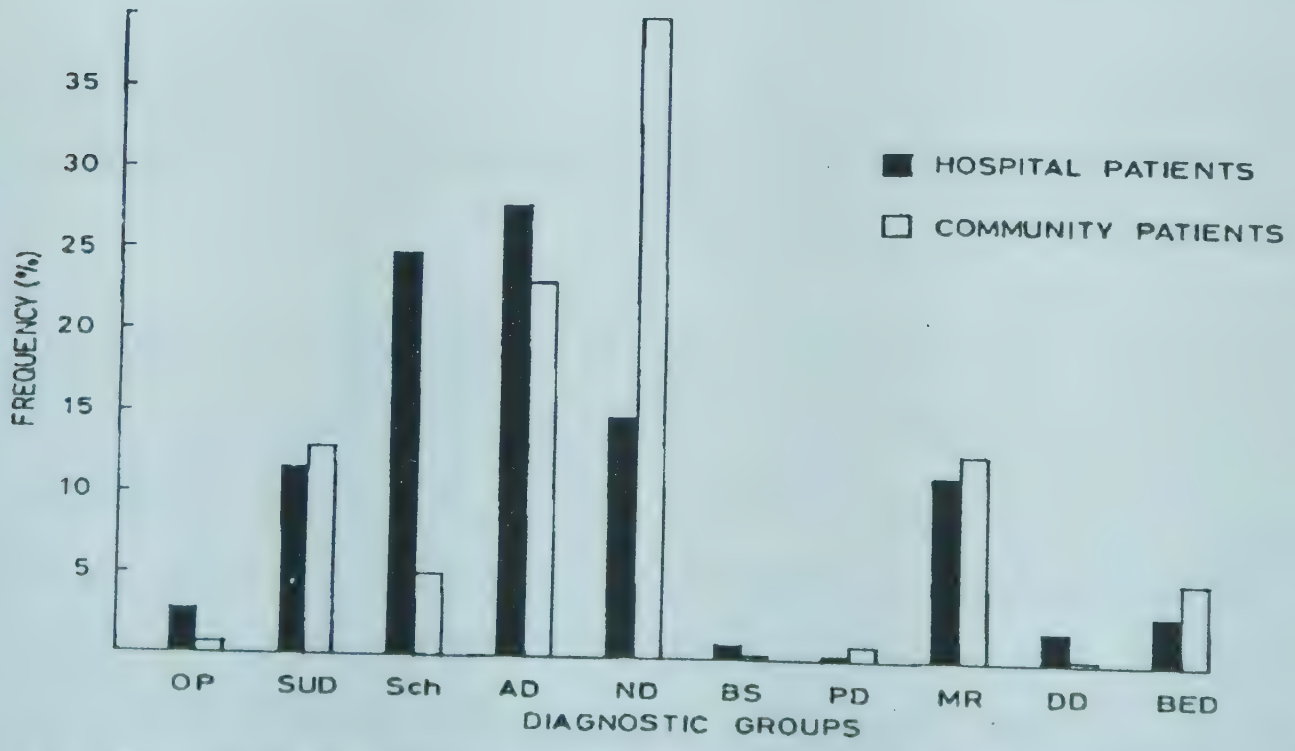


Figure 3.2 Composite Bar Diagram Showing the Diagnostic Distributions of 500 Psychiatric Patients of NIMHANS Hospital and that of Psychiatric Patients in the Indian Community.

In *frequency polygon*, the frequencies are plotted on a graph sheet and the lines joining these points are drawn. A smooth curve is skillfully sketched through these points to obtain frequency curve. More than one frequency polygon may be drawn on the same graph sheet and thus several frequency distributions may be compared. This graphical representation is more appropriate for continuous data as shown in Figure 3.3. It can be viewed that the psychiatric patients of young age groups and old age groups in the community are under represented at NIMHANS hospital.

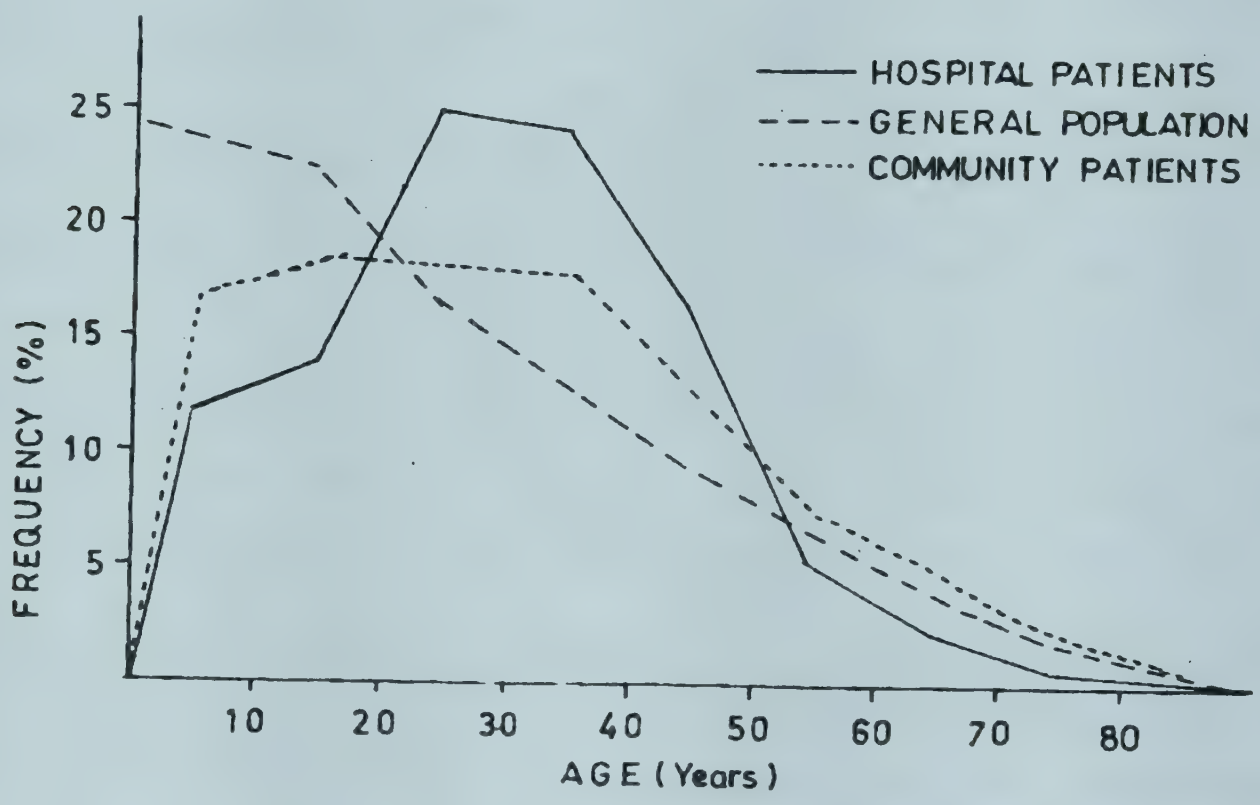


Figure 3.3 Frequency Polygons Showing the Age Distributions of 500 Psychiatric Patients of NIMHANS Hospital, General Population of India and that of Psychiatric Patients in the Community.

In *histogram*, on a common base line, rectangles are drawn with their areas proportional to the frequencies. These rectangles are in juxtaposition and there will be no gaps between them. This diagram is more appropriate for data in ratio level of measurement and especially when the frequency classes are of unequal widths as shown in Figure 3.4. In order to draw this diagram, the frequencies of the classes of varying duration of illness had to be converted into frequencies for unit duration of illness. For

example, the frequency in the frequency class of one month duration is multiplied by twelve in order to obtain the frequency per one year duration. It can be viewed that the load of patients decreases as the duration of illness increases. This point is not clear from the Table 3.1.

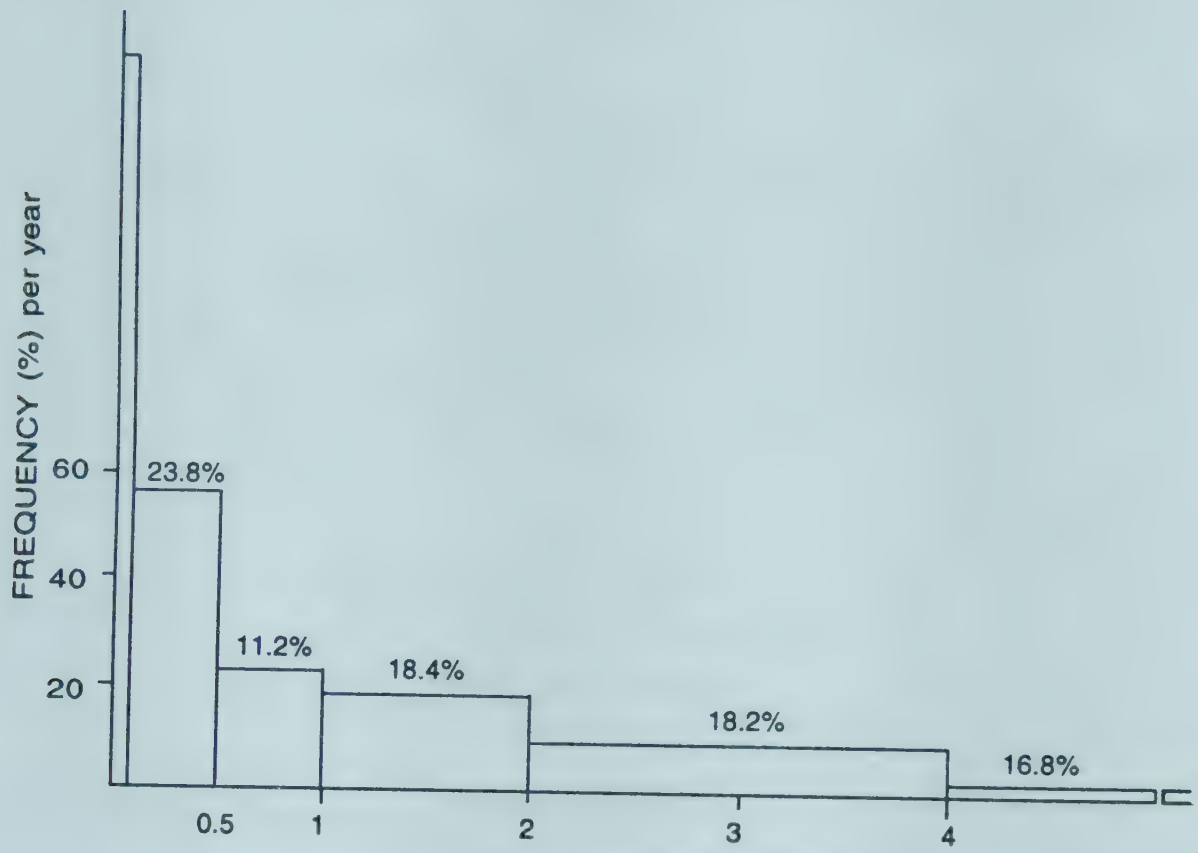


Figure 3.4 Histogram Showing the Duration of Illness of 500 Psychiatric Patients of NIMHANS Hospital.

The geographical data are sometimes arranged alphabetically as shown in Table 3.2. The *spatial character* of the data when presented in tables/graphs is largely obscured. The maps provide the necessary medium for presenting real relationships of spatial data clearly, meaningfully, and adequately. Maps are often indispensable in locating problems, analyzing data and discover hidden facts and relationships. It is advantageous to incorporate both the geographical and political aspects in maps. The geographical aspect is important to determine the causal factors of disorders. The political aspect is important for planning effective mental health delivery system. There are spot maps, shaded maps and graphs superimposed maps. In *spot maps,* the size of the symbol may be proportionate to the magnitude of the

Table 3.2 Government Mental Hospitals in India (N=36) and their Bed Occupancies (Patients) as on First July 1999.

States	Hospitals at	Bed Occupancy
Andhra Pradesh:	Hyderabad (H)	386
	Vishakapatnam (V)	300
Assam:	Tezpur (T)	353
Bihar:	Ranchi (R)	543
	Ranchi CIP (R)	360
Delhi (UT):	Delhi (D)	140
Goa:	Panaji (P)	150
Gujarath:	Ahmedabad (A)	402
	Baroda (B)	181
	Jamnagar (J)	55
	Bhuj (BJ)	25
Jammu & Kashmir:	Srinagar (S)	100
Karnataka:	Bangalore (B)	364
	Dharwad (D)	296
Kerala:	Trivandrum (TM)	774
	Trissur (TR)	382
	Kozhikhode (K)	685
Madhya Pradesh:	Gwalior (G)	192
	Indore (I)	157
Maharashtra:	Thane (T)	1744
	Pune (P)	2540
	Nagpur (N)	786
	Ratnagiri (R)	183
Nagaland:	Kohima (K)	21
Punjab:	Amritsar (A)	415
Rajasthan:	Jaipur (J)	312
Tamil Nadu:	Chennai (C)	1657
Uttar Pradesh:	Agra (A)	459
	Varanasi (V)	258
	Bareilly (B)	292
West Bengal:	Kolkata CPH (K)	251
	Kolkata LPMH (K)	129
	Kolkata IP (K)	36
	Mankundu (K)	106
	Berhampore (B)	214
	Purulia (P)	97

phenomena represented. Symbols may be in the form of circles. The area of the circle may be proportional to the value represented. The spot map for the data in Table 3.2 is as shown in Figure 3.5. The geographical areas which are too far to a government mental hospital in India may be located from this spot map.

Figure 3.5 Spot Map Portraying Bed Occupancies in Government Mental Hospitals in India as on First July 1999

(c) Summarization Figures

The summarization figures of measurement of characteristics of frequency distribution will be helpful to understand the data more clearly. The characteristics for *univariate data* are the average (central tendency), dispersion, skewness and kurtosis.

(i) Average

The observed values of a biosocial variable are not equal, but we can notice a general tendency of these values to cluster around a particular value. It is convenient to characterize and represent each group of observations by such a value which is called as the central tendency or the average of that group. The mean, median and mode are the commonly used measures of average. The *mean* (arithmetic mean) is appropriate when the variable is measured in interval scale. It is obtained by dividing the sum of observations by the total number of observations. The population mean is denoted by μ and that of the sample by $\bar{x}$

Thus,

$$\bar{x} = \frac{1}{n} \Sigma x_i$$

where x_i^s are the observed values and n is the number of observations in the sample. For example, the mean of 4,2,3,0 and 6 is 3.The mean is sensitive to the presence of extreme values in the data, and hence it does not give fair idea of central tendency when the extreme values or outliers are present in the data. Both the change of origin and scale have effect on the mean. The sum of deviations from the mean is always zero and the sum of squares of deviations from the mean is always the minimum.

The *median* is the middle most value of the observations arranged in an ascending or a descending order of magnitude. The ascending

order of 2,1,4,0 and 69 observations is 0,1,2,4,69 and hence the median is 2. It divides the distribution into two equal parts, and hence it is a position value. The other position values are the quartiles, deciles and percentiles. The quartiles divide the distribution into four equal parts, the deciles into ten equal parts and the percentiles divide the distribution into hundred equal parts. The median is appropriate when the variable is measured at least in an ordinal level. The presence of extreme values or outliers have least effect on the median, and hence it gives a good idea of average in case of duration of stay of discharged patients in a mental hospital. The sum of absolute deviations from the median is always the minimum.

The *mode* is that value of the observations which occurs most frequently. For example, the mode for 4,2,3,3 and 6 is 3. There may be any number of modes in a distribution. Thus, we talk of nil modal, unimodal, bimodal, trimodal and in general multimodal distributions. It is the most appropriate measure of average for nominal data. As in the case of the median, the presence of extreme values have least effect on the mode. The median always lies between the mode and the mean. Several empirical relationships between these measures of average have been established such as,

$$(\text{Mean}-\text{Mode}) = 3\ (\text{Mean} - \text{Median})$$

The *proportions and percentages* are the commonly used summarization figures in dealing with categories of qualitative variables.

$$\text{Percentage} = \text{Proportion} \times 100$$

Thus, the percentage frequencies of a categorical variable add to 100 and the proportion frequencies add to 1.

The *rates and ratios* are used as summarization figures when time is one of the components. Hence, they are particularly important

in psychiatric epidemiological research. The results obtained on qualitative variable are expressed as a proportion, percentage, rate or a ratio. A rate measures diseases, disabilities and injuries per unit population. That is,

$$\text{Rate} = \frac{\text{Numerator} \times \text{Time specification} \times K}{\text{Denominator}}$$

$$= \frac{\text{Number infected} \times \text{Time specification} \times 1000}{\text{Population at risk}}$$

When the denominator consists of all those who are capable of getting the disease, it is known as the population at risk.

The rates may be crude, specific and standardized. The *crude rates* are the general rates calculated without paying regard to the specific section of the population such as the crude birth rate. They measure the proportion of the total events to the total population. For more detailed information of the community, the rates may be calculated for specific population groups. Thus we speak of age *specific rates,* gender specific rates, disease specific rates etc. Because of the differences in the structure of the population due to age and gender distributions, the crude rates reveal a rough idea in comparing the health of populations. Only the adjusted rates, called *standardized rates,* adjusted with respect to age and sex distributions give the correct ideas in making comparisons of health of populations or communities. In order to standardize a crude death rate , a standard million population is taken such as the country as a whole. To this population, age and sex-wise death rates of various groups of the population of a place whose crude death rate is to be standardized are to be applied. The weighted average death rate of all age and sex groups of standard million gives the standardized death rate of the place in question. By this method, death rates of any two places can be compared.

The rate should be distinguished from the *ratio.* The numerator comes from the denominator in a rate, whereas the numerator and the denominator are independent in a ratio. Generally, a rate refers to a period of time, whereas a ratio refers to a point of time. For example, the sex ratio (number of females per one thousand males) of 933 of India as on the first March 2001, the number of females (49,57,38,169) is independent of the number of males (53,12,77,078).

(ii) Dispersion

It is only because of variability that we compute averages. We never speak of the average number of days in a week. The observed values of the variate tend to spread over an interval rather than cluster closely around the central average. The variability may be biological, real or experimental. The characteristics of the scatter or spread of the observed values in the neighborhood of the central average is called dispersion. The range, standard deviation and the coefficient of variation are the commonly used measures of dispersion. The *range* is the interval between the maximum and the minimum values of the observations. That is,

$$\text{Range} = X_{max} - X_{min}$$

The range for 4,2,3,0 and 6, is 0 to 6. The range may be applied for the data measured in nominal scale. It depends on only the extreme values and it does not deal with the variations within the group. The range is the crudest measure of dispersion. Its applications are limited to only a few fields. The daily temperature of a place may be studied by noting down the maximum and the minimum temperatures of the place.

The *standard deviation* is the commonly used measure of dispersion when the variable is measured at least in an interval level of

measurement. The standard deviation is the root mean squared deviation measured from the mean. The population standard deviation is denoted by σ and that of the sample by S. Thus,

$$S = \sqrt{\frac{1}{n} \sum (x - \bar{x})^2} = \sqrt{\frac{1}{n} \sum x^2 - \bar{x}^2}$$

Let us suppose that the duration of stay of five patients who were discharged on a particular day in a ward are given by 7, 4, 1, 3 and 5 days. It may be calculated that the mean and standard deviation are 4 days and 2 days respectively. The change of origin has no effect on standard deviation. For example, the standard deviation of 17, 14, 11, 13, and 15 days is the same as those of 7, 4, 1, 3 and 5 days. The change of scale has effect. For example the standard deviation of 70, 40, 10, 30 and 50 days is 2×10=20. The standard deviation is always the minimum among the root mean squared deviation measured from any other average. The square of the standard deviation is the *variance* which may also be used as a measure of dispersion.

The above cited absolute measures of dispersion give risk in comparing the dispersions of two or more distributions since they depend on both the unit of measurement and the average from which the deviations are measured. These measures of dispersion are significant only in relation to the average from which the deviations are measured. The *coefficient of variation* (CV) is a relative measure of dispersion. It is the ratio of the standard deviation to the mean. Generally, it is expressed as percentage. That is,

$$\text{Coefficient of Variation} = \frac{\text{Standard Deviation} \times 100}{\text{Mean}}$$

The coefficient of variation is a pure number. Usually, it lies below 100. Let us suppose that the duration of stay of a group of five

patients are given by 7, 4, 1, 3 and 5 days, and the duration of stay of another group of five patients are given by 4, 2, 3, 0 and 6 days. It may be calculated that,

$$\text{CV for the first group} = \left(\frac{2}{4}\right) \times 100 = 50\%$$

$$\text{CV for the second group} = \left(\frac{2}{3}\right) \times 100 = 67\%$$

Though the standard deviations are the same, the CV of the second group is higher than that of the first group.

(iii) Skewness, Kurtosis

The frequency distributions differ not only with respect to the average and variability, but also with respect to the shape. The characteristics are skewness and kurtosis. The *skewness* is the lack of symmetry. A perfectly bell shaped curve has no skewness. Hence, a distribution is skew if the mean, median and mode do not have the same value. The skewness may be positive such as age distribution of a sample of dissociative (conversion) disorder patients (CON) at NIMHANS hospital, negative such as age distribution of somatoform disorder patients (SOM), or zero such as age distribution of obsessive and compulsive disorder patients (OCD), as shown in Figure 3.6.

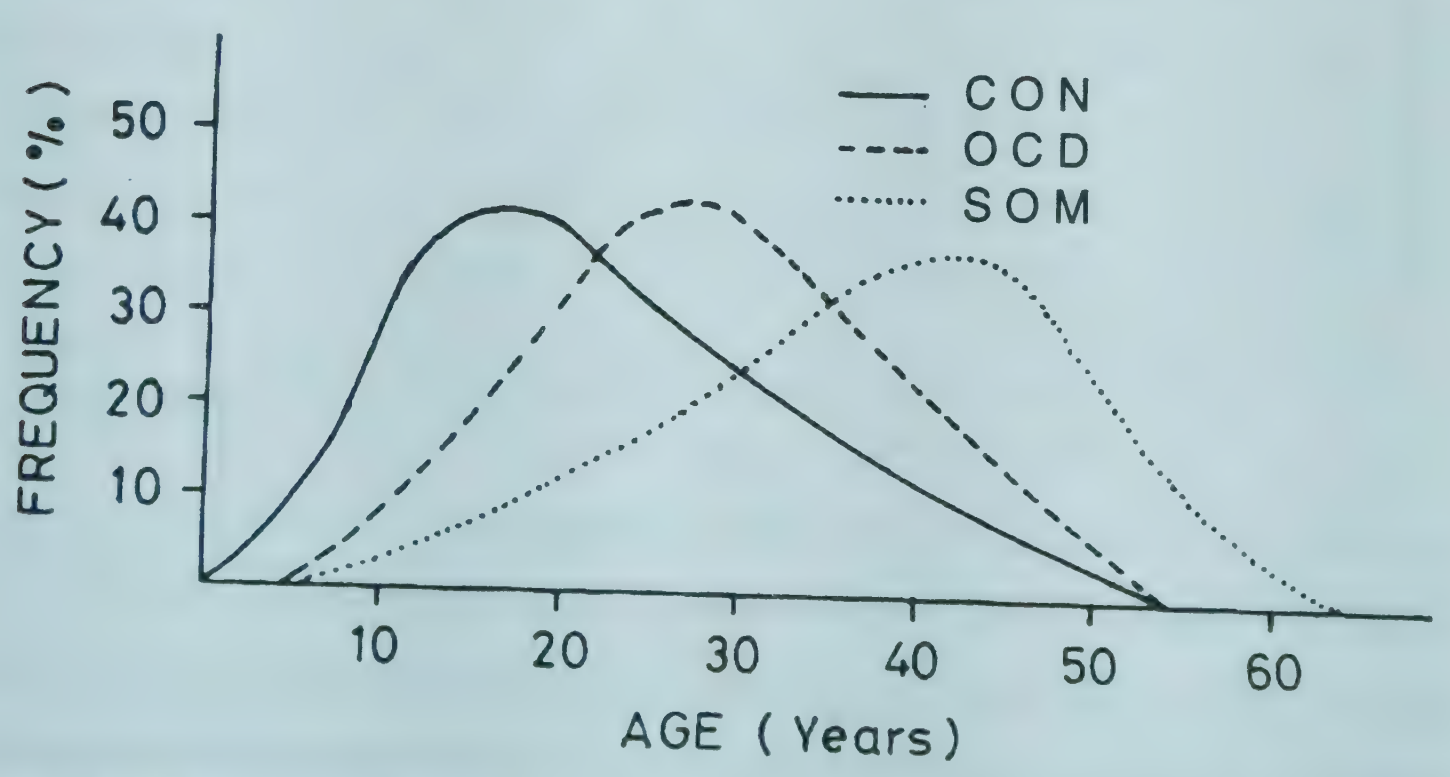


Figure 3.6: Frequency Curves Showing Positively Skewed, Negatively Skewed and Symmetrical Distributions.

One of the measures of skewness which usually lies between -1 and 1, is given by

$$\text{Skewness} = \frac{\text{(Mean-Mode)}}{\text{Standard Deviation}}$$

The kurtosis is the measure of the relative flatness of the top of the frequency curve. A frequency curve may be symmetrical, but may fail to have a peakedness as that of a normal curve. The frequency distributions may have the same variability, but they may be more or less peakedness than that of the normal curve. The kurtosis may be classified as leptokurtic such as the age distribution of obsessional thought disorder (OTD) of a sample of patients at NIMHANS hospital, mesokurtic such as age distribution of obsessive compulsive disorder (OCD) patients, and platykurtic such as the age distribution of the combined group of dissociative disorder patients and somatoform disorder patients (DSD), as shown in Figure 3.7.

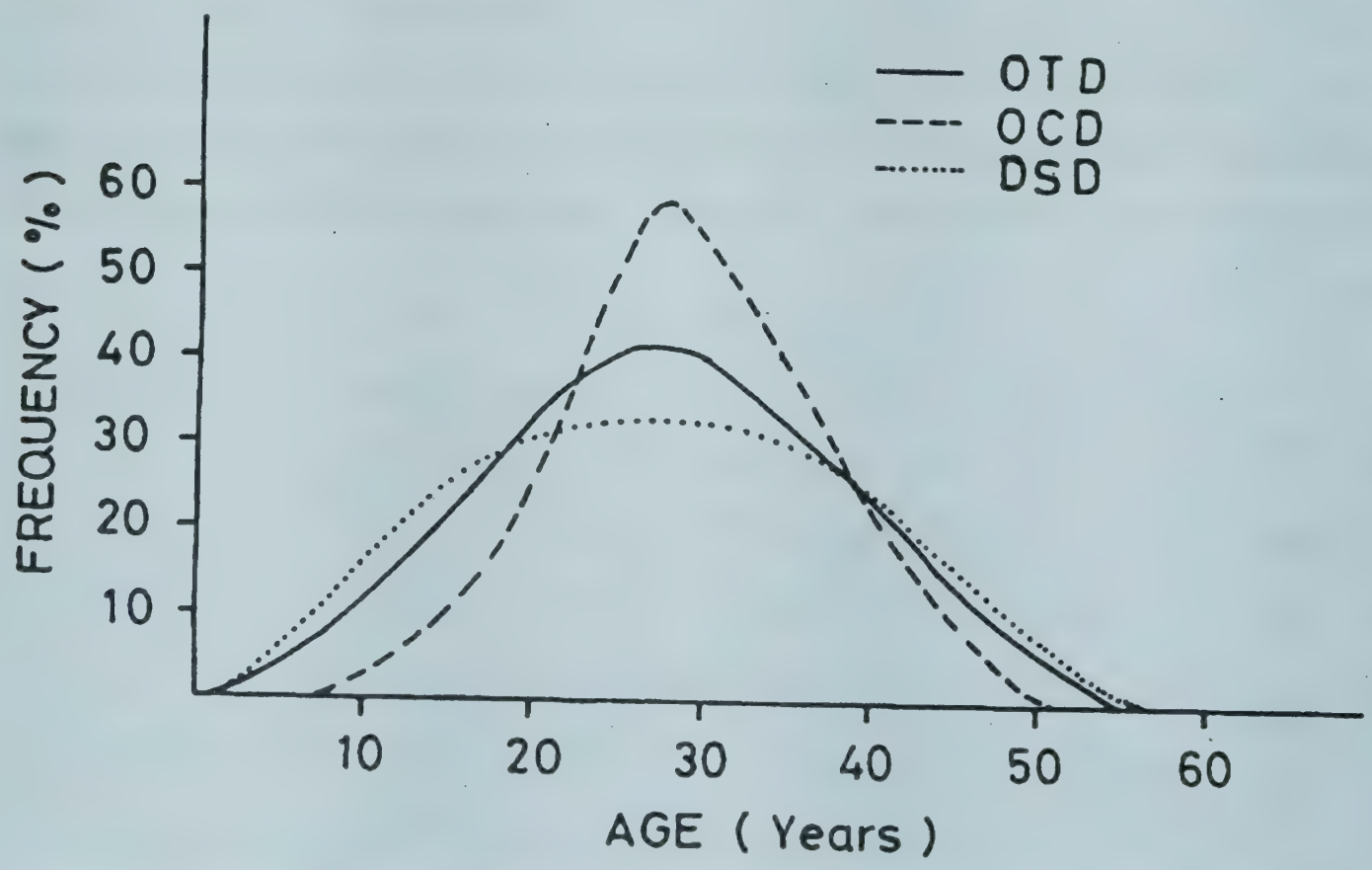


Figure 3.7: Frequency Curves Showing Leptokurtic, Mesokurtic and Platykurtic Distributions.

The kurtosis may be measured by the ratio of the fourth order moment to the square of the variance. This value is above three for leptokurtic, equal to three for mesokurtic (normal distributions) and below three for platykurtic distributions.

CHAPTER 4

POPULATION STATISTICS AND MENTAL HEALTH DELIVERY SYSTEMS IN INDIA

A good mental health information system is a prerequisite for scientific planning and effective administration of mental health services.

(a) Demographic Indicators

Demography is the study of the growth, composition and distribution of human population, and their interrelationships with social, economic and behavioural factors. In addition to statistics relating to existing mental health conditions and available mental health care facilities, demographic indicators are important for specification of goals and targets in terms of measurable outputs. They are used in planning, monitoring and evaluation of mental health programs. The decennial census is the main source of data for demographic studies in India. The registration and survey data are also used as valuable sources of demography.

The *population size* is an important figure as it is utilized in deriving various mental morbidity indices. The population of India was about 1.03 billion (103 crores or 1030 millions) as per the 2001 census. This developing country, which is undergoing demographic transition, has an annual growth rate of about 2%. Being the second most populous country, India is the home for about 16.2 % of the World

population of about 6.3 billions. The country measures a total area of about 33 lakhs square kilometers, accounts for only 2.4% of the World land area of about 15 crores square kilometers. The density of population per square kilometer of a given geographical region is given by,

$$\text{Density of population} = \frac{\text{Total population}}{\text{Area in square kilometers}}$$

The density of population of India may be estimated at 311 persons per square kilometer, while the density of population of the World is estimated at 42 persons during the year 2001. West Bengal is the most densely populated state with 904 persons living per square kilometer. The wide variation in the density of population in various states and union territories of India, as shown in Table 4.1, posses more problems from the point of view of mental health care to the entire population of the country. The density of population in urban localities may be calculated as persons per room to determine overcrowding.

The *age distribution* of the population provides the true picture regarding the risks of various mental and behavioural disorders associated with different age groups. The age distribution of Indian population clearly indicates that ours is an young population composing about 42% in the age group of 0-14 years. The aged people of sixty years and above constitute only about 6.5 % of the population. The *sex distribution* of the population will help in studying different mental and behavioural disorders associated with each sex. This is important for improving the mental health care of the entire population. The sex ratio of Indian population has shown a declining trend from the beginning of the century and it was about 933 during the last census year. The number of females is always higher than

Table 4.1 Population, Area and Density of Population of States/Union Territories of India -1991

States/Union Territories	Population (crores)	Area (lakhs of sq. kms)	Density
Southern Region			
1. Andhra Pradesh	6.7	2.8	242
2. Karnataka	4.5	1.9	234
3. Kerala	3.0	0.4	747
4. Tamil Nadu	5.6	1.3	428
Northern Region			
5. Haryana	1.6	0.4	372
6. Himachal Pradesh	0.5	0.6	92
7. Jammu & Kashmir	0.8	2.2	35
8. Punjab	2.0	0.5	401
9. Rajasthan	4.4	3.4	129
10. Uttaranchal*	0.7	0.6	126
Central Region			
11. Madhya Pradesh	4.9	3.1	158
12. Uttar Pradesh	13.2	2.4	554
Western Region			
13. Goa	0.1	0.04	316
14. Gujarat	4.1	2.0	210
15. Maharastra	7.9	3.1	256
Eastern Region			
16. Arunachal Pradesh	0.09	0.8	10
17. Assam	2.2	0.8	284
18. Bihar	6.5	0.9	686
19. Chhatisgarh*	1.8	1.4	130
20. Jharkhand*	2.2	0.8	273
21 . Manipur	0.2	0.2	82
22. Meghalaya	0.2	0.2	78
23. Mizoram	0.1	0.2	33
24. Nagaland	0.1	0.2	73
25. Orissa	3.2	1.6	202
26. Sikkim	0.04	0.07	57
27. Tripura	0.3	0.1	263
28. West Bengal	6.9	0.9	766
Union Territories			
1. Andaman & Nicobar Islands	0.03	0.08	34
2. Chandigarh	0.06	0.001	5631
3. Dadra & Nagar Haveli	0.01	0.005	282
4. Daman & Diu	0.01	0.001	906
5. Delhi	0.9	0.01	6319
6. Lakshadweep Islands	0.005	0.0003	1615
7. Pondicherry	0.08	0.005	1605
Total	85.	33.	258

* new states formed in 2000

males in socio-economically advanced countries, while in the under-developed countries, the number of males is higher than females due to better health care of males. The domicile characteristic (distinction between rural and urban localities) of the population will help in preparing the required mental health programs for these sectors separately. The percentage of urban population of India has shown an increasing trend from the beginning of the century and it was about 30% during the last census year.

(i) Vital Statistics

The statistics of vital events such as births and deaths are known as vital statistics. The civil registration scheme, sample registration scheme and model registration scheme provide the basic data for vital statistics of India. A knowledge of the *rates of fertility* in a community is important for planning mental health facilities and services for mothers and others. The annual crude birth rate (CBR) of a geographical area is the general rate of fertility. It is defined as,

$$\text{CBR} = \frac{\text{Number of live births during the year} \times 1000}{\text{Population as on the first July of the year}}$$

The CBR of India has shown a declining trend from the beginning of the century and it was about 29.5 during the 1991 census year. The annual general fertility rate (GFR) of a geographical area is a specific rate of fertility. It is defined as,

$$\text{GFR} = \frac{\text{Number of live births during the year} \times 1000}{\text{Number of women in their child bearing age}}$$

The child bearing age usually goes from 15 to 49 years. The attractive feature of this rate is that here the denominator approximates the

number of persons actually exposed to the risk of bearing a child. The GFR of India was about 124 during the 1991 census year.

The *high mortality* rates are related to the low living, and to the lack of medical and health facilities. The annual crude death rate (CDR) of a geographical area is the general rate of mortality. It is defined as,

$$\text{CDR} = \frac{\text{Number of deaths during the year} \times 1000}{\text{Population as on the first July of the year}}$$

The CDR of India has shown a decreasing trend from the beginning of the century and it was 9.8 during the 1991 census year. The annual infant mortality rate (IMR) of a geographical area is a specific rate of mortality. It is defined as,

$$\text{IMR} = \frac{\text{Number of deaths of babies aged below one year during the year} \times 1000}{\text{Number of live births during the year}}$$

This is an important rate of mortality since it directly measures the health of mothers, infants and children, and further it is associated with several social and economic factors. The IMR of India was 80 during the 1991 census year. It was fixed at less than 60 for health by 2000 AD.

(ii) Measures of Mental Morbidity

The incidence rate and the prevalence rates are the important indices to measure epidemiological situation, to suggest priorities and to asses the progress made in the control of disorders. The incidence rate is the rate at which people without the disease develop the disease during a specified period of time. That is,

$$\text{Incidence Rate} = \frac{\text{Number of new cases during the period} \times 1000}{\text{Population at risk during the period}}$$

The incidence rates are usually determined for acute diseases as their onsets are sharply defined. However these rates are important to answer question whether the growth of urbanization is leading to increase in mental stress and hence mental disorders.

The point prevalence rate is the rate at which people having the disease at a particular point of time. That is,

$$\text{Point Prevalence Rate} = \frac{\text{Number of cases at a particular point of time} \times 1000}{\text{Population at risk at the particular point of time}}$$

It is not possible to survey the entire defined population at a point of time to arrive at the point prevalence of a disease. The procedure to be followed is to continue the survey till completed, not taking into consideration any occurrences reported in the already surveyed population. The period prevalence rate is the rate at which people having the disease during a specified period of time. That is,

$$\text{Period Prevalence Rate} = \frac{\text{Number of cases (new \& old) during a specified period} \times 1000}{\text{Population at risk during the period}}$$

It is obvious that the period prevalence rate is the sum of the point prevalence rate at the beginning of the period and the incidence rate. The prevalence rates are determined for chronic illnesses such as mental and behavioural disorders. The lack of absolute standard criteria for defining a case is the foremost among the methodological problems in psychiatric epidemiology field. A meta-analysis (techniques used for the systematic synthesis of results from many studies) of mental morbidity studies carried out in India, yielded a prevalence rate of 58.2 per one thousand population. The details are as shown in Table 4.2.

Table 4.2 Prevalence Rates (Per one thousand Population) of Mental and Behavioural Disorders in India

Diagnostic Blocks	Prevalence Rate
Organic Psychoses	0.4
Substance use Disorders	6.9
Schizophrenia	2.7
Affective Disorders	12.3
Neurotic Disorders	21.3
Personality Disorders	0.6
Mental Retardation	6.9
Behaviour/Emotional Disorders	2.7
Epilepsy	4.4
Total	58.2

The affective disorders (12.3 prevalence rate) include mania (0.7), manic depression (2.7) and psychotic depression (8.9). The neurotic disorders (21.3) include phobias (4.2), other anxiety disorders (5.8), neurotic depression (3.1), obsessive-compulsive disorders (3.1), dissociative/conversion disorders (4.5) and somatoform disorders (0.6). Based on this rate, it can be estimated that there were 5.97 crores psychiatric patients in India as on first March 2001. Further analysis revealed that the prevalence rate for females (64.8) is significantly more than those of males (51.9). Only the prevalence rates of mania, mental retardation and substance use disorders are significantly more among males. Based on these rates it can be estimated that there were 2.76 crores male psychiatric patients and 3.21 crores female psychiatric patients in India as on first March 2001. Thus the number of female psychiatric patients are more (53.8%) than the number of male psychiatric patients (46.2) in India. The prevalence rate for urban people (80.6) is significantly more than those of rural people (48.9). Only the prevalence rates of organic psychoses, epilepsy,

hysteria and somatic complaints are significantly more among rural people.

(b) Mental Hospital Service Indicators

The mental hospital statistics have several administrative and clinical values. Dr. S.M.Channabasavanna and Dr. S.D.Sharma are pioneering in utilizing these statistics for the development of effective mental health delivery systems for India. The disagreement in diagnostic categorization is the major methodological problem in the maintenance of mental hospital statistics. The methods used in clinical psychiatry, namely conversation and observation are rather crude. Unlike in general medicine, the diagnostic categories in psychiatry are based on symptomatology and not on aetiology which is usually multifactorial. Frequently there is a substantial overlap in the symptom content of various psychiatric syndromes. The characteristics of hospital patients are based on the disorders for which the hospital facilities are being sought and hence they may not be a fair representation of the disorders in the surrounding community. Though hospital data do not provide the estimate at the community level, they are useful in spelling out the dimensions of the problems and their applications in planning process and quality of life ascertainment. The annual and seasonal variations of the disorders in the community may be reflected by changes in the new registrations of the hospital. Eighteen indicators measuring various aspects of mental hospital facilities and services are defined for reliable evaluation and comparisons of these hospitals with regard to their programs and quality of service. These indicators and their trends are presented in Table 4.3 for government mental hospitals in India for the sake of information of the readers.

(i) Out-Patients

The out-patient service forms an integral part of the total service rendered by a mental hospital. The number of new registrations during

the year and the follow-up rate are the important indicators. The latter is defined as,

$$\text{Follow-up Rate} = \frac{\text{Follow-up attendance during the year}}{\text{Number of new registrations during the year}}$$

(ii) Bed Strength, In-patients

The bed strength of the hospital indicates the maximum number of patients who could be institutionalized at a given point of time. It is sanctioned by the controlling authority based on the accommodation, necessary facilities, man-power and budget. On the other hand, the average bed occupancy is the average number of inpatients staying in the hospital. It is given by,

$$\text{Average bed occupancy} = \frac{\text{Hospital days of the year}}{365}$$

The hospital days of the year is obtained by adding the hospital in-patient census figures of every day over the year. The denominator is 366 in case of leap year. The average bed occupancy of the hospital corresponds to the point prevalence rate of the community. The over-crowding of the hospital may be measured by,

$$\text{Over crowding (\%)} = \frac{(\text{Average bed occupancy} - \text{Bed strength}) \times 100}{\text{Bed strength}}$$

The *characteristics of in-patients* on a particular day, say first July of the year, is an important aspect to compare and evaluate the services of hospitals. The important characteristics of in-patients in a mental hospital are: (1) duration of stay, (2) age, (3) gender, (4) mode of admission, and (5) diagnostic groups.

(iii) Rate of Turnover, Discharged Patients

The rate of turnover yields the number of patients who were treated as in-patients in the hospital during a given period of time, say one year. Thus, the number of discharged patients during a particular year is an absolute indicator. The average duration of stay (ADS) of hospitalized patients is an important indicator of the rate of turnover. But the evaluator has to take a decision as to which type of patients the ADS is to be calculated. The ADS has to be the number of days that a patient can be expected to stay in the hospital at the time of admission to in-patient treatment. The ADS may be calculated based on in-patients on a particular day, or it may be calculated based on discharged patients during a period of time. The ADS based on in-patients may not be preferable since these patients could not complete their stay in the hospital at the time of calculating the average. Further, this ADS is unstable as it is unduly effected by the load of chronic patients. The ADS based on discharged patients could not take into consideration of the stay of all the patients stayed during the period. The simplest method of calculating the ADS in a mental hospital is the one that is based on the hospital days of the year. It is defined as,

$$\text{ADS (hospital days)} = \frac{\text{Hospital days of the year}}{\text{Number of discharged patients during the year}}$$

This indicator tends to mean duration of stay as the load of chronic patients decrease, and tends to median duration of stay as the load of chronic patients increase.

The *characteristics of discharged patients* is another aspect to compare and evaluate the services rendered by the hospital. The important

characteristics are the distribution according to results of treatment, and the hospital death rate (HDR). The latter is defined as,

$$\text{HDR} = \frac{\text{Number of deaths during the year} \times 1000}{\text{Average bed occupancy during the year}}$$

Thus, the hospital death rate of hospital is defined analogous to the crude death rate of the general population to facilitate comparisons. It is advantageous to standardize the HDR according to age and gender distributions of in-patients for drawing more meaningful conclusions on the mortality rates of psychiatric patients.

(iv) Man-power, Expenditure Pattern

Adequate man-power is required to provide services to the hospital patients. The number of psychiatrists, and the average number of in-patients per psychiatrist are important indicators. The latter is defined as,

$$\text{In-patients per psychiatrist} = \frac{\text{Average bed occupancy}}{\text{Number of psychiatrists}}$$

Sufficient budget has to be sanctioned to provide adequate facilities and services to the hospital patients. The total expenditure and the unit cost (cost per day per patient) for a given financial year are the important indicators. The latter is defined as,

$$\text{Unit Cost} = \frac{\text{Total expenditure on service during the financial year}}{\text{Hospital days of the financial year}}$$

Table 4.3 Service Indicators and their Trends of Government Mental Hospitals in India (N=36)

Area of Service/Indicators	For the Year 1993	Significant Trends (1977-93)
I Out-Patients:		
1. Average number of registrations per hospital	5742 patients	Increasing
2. Follow-up rate	3.1	Increasing
II. Bed Strength, In-patients:		
3. Average bed strength per hospital	602 beds	Decreasing
4. Average bed occupancy	528 patients	Decreasing
5. Over-crowding	-12%	Decreasing
6. Duration of stay: Above 2 years	51%	Decreasing
7. Age : Above 60 years	8%	No trend
8. Gender : Males	61%	No trend
9. Mode of admission		
Voluntary	48%	No trend
Certified	42%	No trend
Observations	7%	No trend
Criminals	3%	No trend
10. Diagnoses: Psychoses	76%	Decreasing
III. Rate of Turnover, Discharged Patients:		
11. Average number of discharged per hospital	1667 Patients	Increasing
12. Average duration of stay	112 days	Decreasing
13. Result of treatment:		
Improved	93%	No trend
Not improved	5%	No trend
Deaths	2%	No trend
14. Hospital Death Rate	53.1	No trend
IV. Man-power, Expenditure Pattern:		
15. Average number of psychiatrists per hospital	6.2	Increasing
16. In-patients per psychiatrist	86	Decreasing
17. Average expenditure per hospital - 1996	Rs 2.1 crores	Increasing
18. Units Cost - 1996	Rs 129.00	Increasing

(c) Medical College Hospital Psychiatric Units

The number of Medical College Hospital Psychiatric Units has increased from 125 to 175 during the period from 1988 to 1999, and these units fomed the next major mental health delivery system in India. For the year 1999, the average number of registration per hospital was 3055 patients, while the fallow-up attendence ~~to~~ was about 3.3 times more as shown in the Table 4.4. Both these indicators of outpatient service were in increasing trend during the period from 1988 to 1999. Similarly, the age, sex and diagnostic distributions and the trends are also shown in the table. During the period , the average annual bed occupancy per unit has decreased from 25 beds to 22 beds, while the average number of discharged patients per unit ~~hs~~ has increased from 438 patients to 445 patients, and hence the average duration of stay has decreased from 21 days to 18 days. The average number of psychiatrists per unit has increased from 2.3 to 3.2, while the average number of inpatients per psychiatrist has decreased from 11 patients to 7 patients.

(d) Other Mental Health Delivery Systems

The other mental health delivery systems include the private psychiatric nursing homes and clinics, and psychiatric clinics in district hospitals and in district prisons. The Indian System of medicine of dealing wih psychiatric patients has been active for centuries of years. The majority of rural psychiatric patients visit local agencies / traditional healers such as temple priests, astrologers and mantravadies.

Table 4.4 Service Indicators of Medical College Hospital Psychiatric Units in India for the year 1988 and 1999

Area of Service/Indicators (Number of units)	1988 (125)	1999 (175)
I Out-Patients:		
1. Average number of registrations per Unit (Patients)	2914	3055
2. Age (years) :		
Below 15	5.5%	6.3%
15-60	90.4%	89.0%
Above 60	4.1%	4.7%
3. Gender : Females	40.1%	40.8%
4. Diagnosis : Psychoses	49.8%	45.9%
5. Follow-up rate	3.2	3.3
II. Bed Strength, In-patients:		
6. Average bed strength per unit (beds)	25	22
III. Rate of Turnover, Discharged Patients:		
7. Average number of discharged patients per unit	438	445
8. Average duration of stay (days)	21	18
IV. Man-power		
9. Average number of ~~inpatients~~ Psychiatrists per Unit	2.3	3.2
10. Average number of inpatients per psychiatrist (patients)	11	7

CHAPTER 5

BIVARIATE STATISTICAL METHODS

All statistical methods which simultaneously analyze two variables are known as bivariate statistical methods. The two variables in bivariate data analysis are usually denoted by x and y. The x may represent the reading ability of a group of students in a class and y may represent their spelling ability, x may represent their heights and y may represent their weights etc. Such data presenting two sets of related measures arise frequently in mental health care research. The correlation analysis and the regression analysis are important statistical methods in the bivariate case.

(a) Correlation Analysis

The correlation is the relationship *between variables.* There will be a positive correlation if both the variables increase or decrease simultaneously. The correlation is negative if one variable increases while the other variable decreases. The primary step in correlation analysis is to present the bivariate data on a graph sheet and grasp the type of relationship. Such a graphical representation is known as dot diagram or scatter diagram.

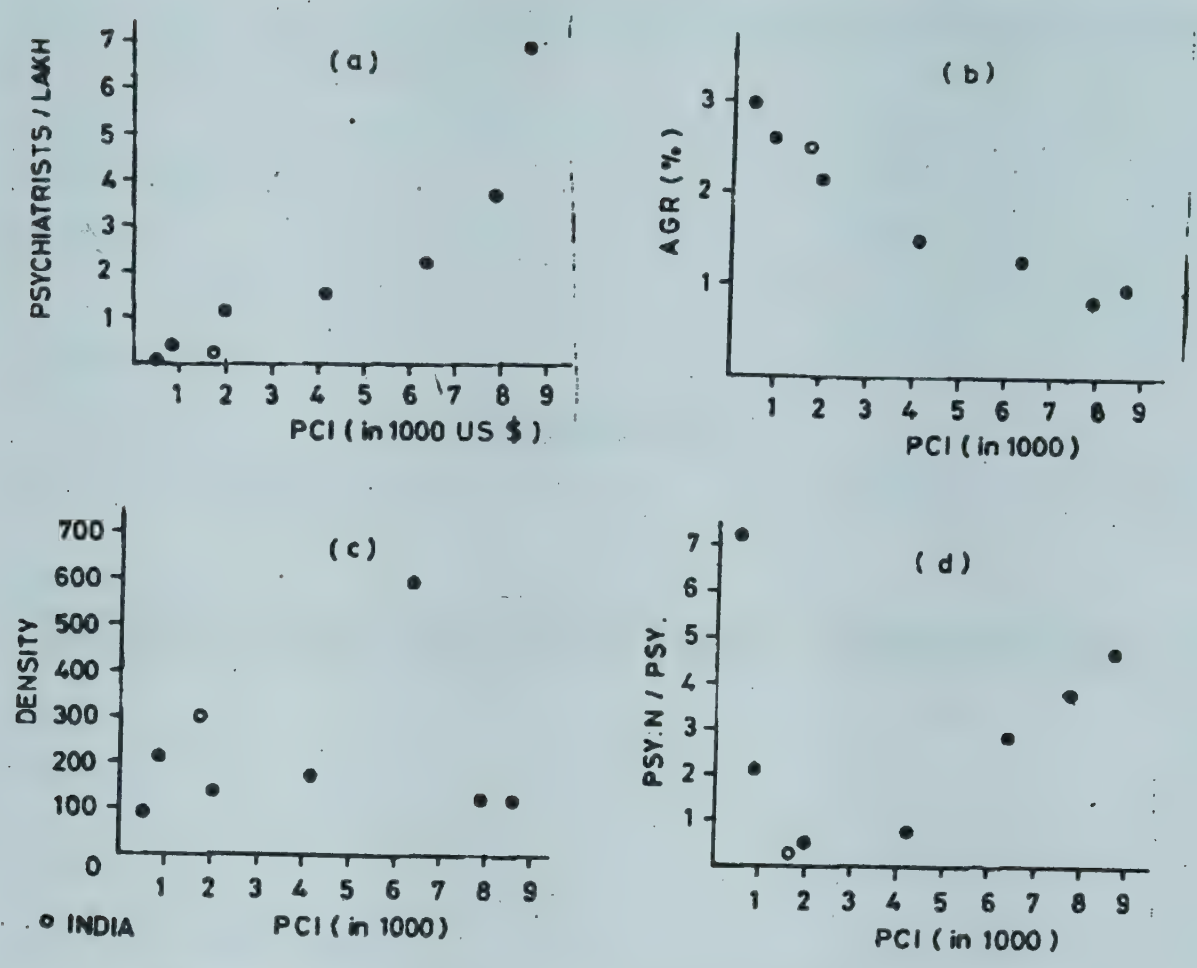


Figure 5.1 Dot Diagrams Showing, (a) Positive Linear Relationship, (b) Negative Linear Relationship, (c)Nil Relationship, and (d) Curvilinear Relationship

The dot diagrams showing the type of relationships between per capita income (PCI) and the number of psychiatrists per one lakh population, between PCI and annual growth rate of population, between PCI and density of population, and between PCI and number of psychiatric nurses per psychiatrist man-power of eight countries are presented in Figure 5.1.

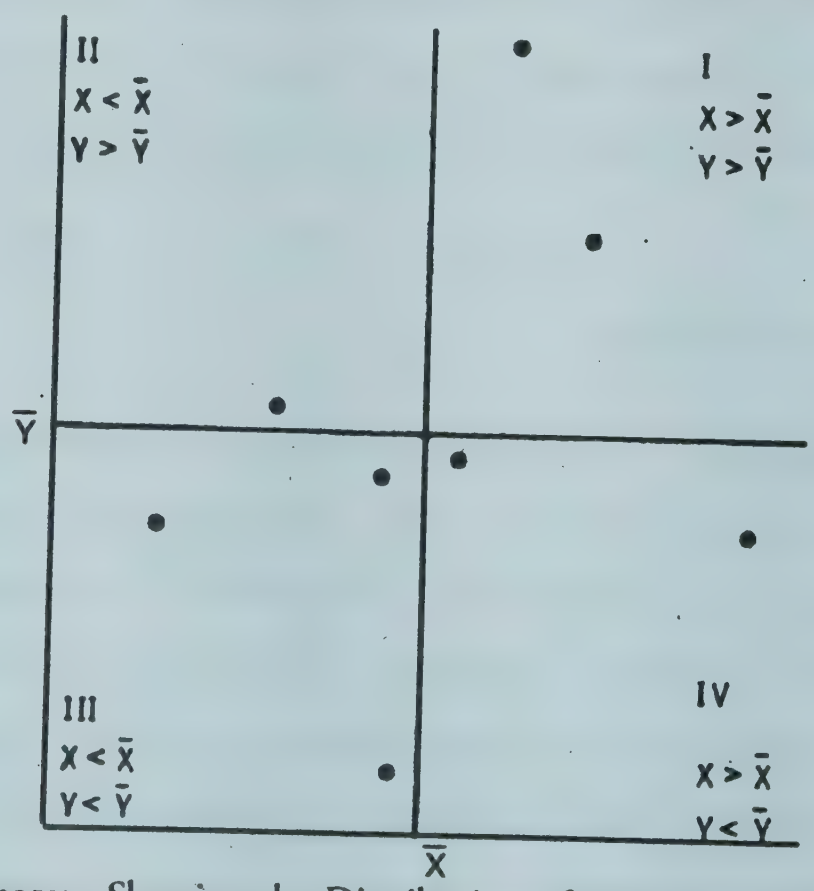


Figure 5.2 A Dot Diagram Showing the Distribution of Dots Over the four Quadrants.

We limit our study to *linear relationship* and the theory built upon this assumption is known as linear or simple correlation. Professor Karl Pearson has given the measurement of this type of relationship. He divided the dot diagram into four parts called quadrants as shown in Figure 5.2. The x and y are the variables in the bivariate data, $\bar{x}$ and $\bar{y}$ are the means of observations of their respective variables, and n is the sample size.

If the trend of the dots is from quadrant III to quadrant I, then there will be a positive linear relationship. If the trend of the dots is from quadrant II to quadrant IV, then there will be a negative linear relationship. Hence, the dots in quadrants I and III indicate positive correlation, and the dots in quadrants II and IV indicate negative correlation. The product $(x - \bar{x}) (y - \bar{y})$ is positive for any dot in quadrants I and III, and it is negative for any dot in quadrants II and IV. Hence, the sum of the products $\Sigma (x - \bar{x}) (y - \bar{y})$ describes the trend of the dots over the four quadrants. Consequently, a natural measure of association would be obtained by dividing it by n. The resultant is known as the covariance of x and y, denoted as Cov (x,y). That is,

$$\text{Cov}(x,y) = \frac{1}{n} \Sigma (x - \bar{x}) (y - \bar{y})$$

The covariance may be made independent of the unit of measurement by dividing it by the product of the standard deviation of x and the standard deviation of y. The resultant is denoted by r. That is,

$$r = \frac{\text{Cov}(x, y)}{\sqrt{\text{Var}(x)\ \text{Var}(y)}}$$

$$= \frac{\Sigma xy - n\bar{x}\bar{y}}{\sqrt{(\Sigma x^2 - n\bar{x}^2)(\Sigma y^2 - n\bar{y}^2)}}$$

This product moment correlation coefficient is known as the pearson correlation coefficient or simply the correlation coefficient.

Let us suppose that the scores on a reading ability test (x) and on a spelling ability test (y) of a group of five students are as given below.

Students	x	y
A	4	7
B	2	4
C	3	1
D	0	3
E	6	5

For this bivariate data, $n = 5$, $\bar{x} = 3$, $\bar{y} = 4$, $\Sigma x^2 = 65$, $\Sigma y^2 = 100$ and $\Sigma xy = 69$. Thus,

$$r = \frac{69 - 5 \times 3 \times 4}{\sqrt{(65 - 5 \times 3^2)\,(100 - 5 \times 4^2)}} = 0.45$$

The assumptions underlying the calculation and use of r are that the two variables are measured in interval scale, normally distributed and linearly related. The r lies between -1 and 1. The $r = -1$ means perfect negative linear relationship, $r = 1$ means perfect positive linear relationship, and $r = 0$ means no linear relationship. The value of r should not be expressed as proportion. For example, $r = 0.5$ does not lie exactly between zero and 1. In fact, r^2 is the coefficient of determination and $(1-r^2)$ is the coefficient of non-determination. Both the change of origin and scale has no effect on the correlation coefficient, suggesting a short-cut method to calculate r. The high correlation coefficient does not indicate that one variable is the cause

and the other is the effect. The r is defined for a specific population, traits and situation.

(i) Specific Measures of Association

The pearson correlation coefficient which was established for bivariate data in which both the variables are measured in interval level, may be logically extended to the lower levels of measurement. These measures are important in mental health care research as most of the variables are measured in nominal and ordinal scales. The *pie correlation coefficient* (ϕ) is computed for the data in which both the variables being correlated are genuine dichotomous and the classes are separated by a real gap between them. It is defined as,

$$\phi = \frac{(ad-bc)}{\sqrt{(a+b)\,(c+d)\,(a+c)\,(b+d)}}$$

where an example for data format is given by,

Gender	In-patients	Out-patients	Total registrations
Males	a	b	(a+b)
Females	c	d	(c+d)
Total	(a+c)	(b+d)	n

The pie correlation coefficient lies between -1 an 1. A classification according to gender and type of service of a representative sample of psychiatric patients of NIMHANS hospital is as presented below.

Gender	In-patients	Out-patients	Total registrations
Males	111 (70.2)	192 (56.1)	303 (60.6)
Females	47 (29.8)	150 (43.9)	197 (39.4)
Total	158 (100.0)	342 (100.0)	500 (100.0)

In this two-way classified data, a = 111, b = 192, c = 47, and d = 150. Thus

$$\phi = \frac{(111 \times 150 - 192 \times 47)}{\sqrt{303 \times 197 \times 158 \times 342}} = 0.134$$

The male gender and in-patient service are positively associated and the phi coefficient is 0.134.

The *contingency coefficient* (CC) is computed from the data in which both the variables are measured in nominal scale. It is defined as,

$$CC = \sqrt{\frac{\chi^2}{n + \chi^2}}$$

where χ^2 (chi-square) is given by,

$$\chi^2 = \Sigma \frac{(o_{ij} - e_{ij})^2}{e_{ij}} = \Sigma \frac{o^2_{ij}}{e_{ij}} - n$$

The o_{ij} are observed frequencies and the e_{ij} are expected frequencies. The expected frequencies are based on the assumption that the variables are independent. Thus the expected frequency of a cell in the contingency table is obtained by dividing the product of its marginal totals by the total of the frequencies. The contingency coefficient lies between zero and less than one. A classification according to religion

and type of service of 500 psychiatric patients of NIMHANS hospital is as presented below.

Gender	In-patients	Out-patients	Total registrations
Hindu	133 (84.2)	273 (79.8)	406 (81.2)
Muslim	14 (8.9)	52 (15.2)	66 (13.2)
Christian	11 (6.9)	17 (5.0)	28 (5.6)
Total	158 (100.0)	342 (100.0)	500 (100.0)

Here, n = 500. For example, the observed frequency of Hindu in-patients is 133 while the expected frequency is calculated to be ,

$$O_{11} = \frac{406 \times 158}{500} = 128.3$$

and thus $$\chi^2 = \left[\frac{133^2}{128.3} + \frac{273^2}{277.7} + - - - - + \frac{17^2}{19.2}\right] - 500 = 4.39$$

and $$CC = \sqrt{\frac{4.39}{500+4.39}} = 0.093$$

The Hindu and Christian religions are positively associated with in-patient service.

The *spearman rank correlation coefficient* (r_s) is derived for the situation in which the observed values of both the variables are expressed as ranks. This rank difference correlation coefficient is given by,

$$r_s = 1 - \frac{6 \Sigma d_i^2}{n (n^2 - 1)}$$

The n is the number of individuals, the di is the difference between the ranks of the observation on the x variable and the ranks of the observation on y variable of the i^{th} individual. This rank correlation coefficient lies between -1 to 1. The ranks of the reading ability test scores and the ranks of the spelling ability test scores of the five students (as given in the illustration of the calculation of r) are as presented below.

Students			Ranks of			
	x	y	x	y	d_i	d_i^2
A	4	7	2	1	1	1
B	2	4	4	3	1	1
C	3	1	3	5	-2	4
D	0	3	5	4	1	1
E	6	5	1	2	-1	1
Total	15	20	15	15	0	8

In this bivariate data, $n = 5$ and $\Sigma d_i^2 = 8$. Hence,

$$r_s = 1 - \frac{6 \times 8}{5\,(5^2 - 1)} = 1 - 0.4 = 0.6$$

The rank correlation coefficient between the reading ability test scores and the spelling ability test scores is 0.6, while the pearson correlation coefficient is 0.45.

The *biserial correlation coefficient* is developed for the situation in which both the variables are measured in interval scales or continuously measured, but one of them is for some reason reduced to two categories. The point biserial correlation coefficient (r_{pb}) is appropriate when one of the two variables is a genuine dichotomy. A special formula is provided which does not resemble the basic pearson formula. It reads,

$$r_{pb} = \frac{(M_p - M_q)\sqrt{pq}}{\sigma_x}$$

where M_p is the mean of the continuous variable for the higher group in the dichotomous variable, M_q is the mean of the continuous variable for the lower group in the dichotomous variable, p is the proportion of the number of cases in the higher group of the dichotomous variable, q = (1– p), and σ_x is the standard deviation of the continuously measured variable for the whole sample. The reading ability test scores (x) and the gender of the five students are presented below.

Students	x	Gender
A	4	male
B	2	male
C	3	female
D	0	female
E	6	male

In this bivariate data, M_p= 4, M_q= 1.5, p= 0.6, q= 0.4 and σx=2. Then,

$$r_{pb} = \frac{(4–1.5)\sqrt{0.6 \times 0.4}}{2} = 0.612$$

The male gender is associated with higher scores and the association is 0.612.

A *tetrachoric correlation coefficient* (r_{tet}) is computed for the data in which both the variables being correlated have been reduced artificially to two categories. The calculation and use of this correlation coefficient are based on the assumptions that the variables are measured in interval scale, normally distributed and linearly related. Under these

assumptions, it gives a coefficient that is numerically approximate to the r. It is the way of estimating the correlation between the two variables when the data could not be obtained in graded quantities. A commonly used formula for tetrachoric correlation coefficient is given by,

$$r_{tet} = \cos\left(\frac{180^0}{1+\sqrt{\frac{ad}{bc}}}\right)$$

Where an example for data format is given by,

Variable 1	Variable 2	
	High	Low
High	a	b
Low	c	d

The tetrachoric correlation coefficient lies between -1 to 1. The reading ability test scores (x) and the spelling ability test scores (y) of the five students (described in the calculation of r) are reduced to two categories by using their means as cutting points, as presented below.

Students			Dichotomous	
	x	y	x	y
A	4	7	High	High
B	2	4	Low	High
C	3	1	High	Low
D	0	3	Low	Low
E	6	5	High	High

For this bivariate data, a = 2, b = 1, c =1, and d = 1. Hence,

$$r_{tet} = \cos\left(\frac{180^0}{1+\sqrt{\frac{2}{1}}}\right) = \cos(74.6) = 0.266$$

The tetrachoric correlation coefficient between the variables is 0.266, while the pearson correlation coefficient is 0.45

(b) Regression Analysis

The main purpose of regression analysis is to predict the value of the dependent variable (y) when the value of the independent variable (x) is given or known. The analysis describes the dependence of a variable on an independent variable, suggest possible cause and effect relationship between factors, and explain some of the variation of the dependent variable by the independent variable by using the latter as a control. Studies such as the effect of urbanization on the increase of prevalence of neurotic disorders of a defined population, age of person on the spiritual dimension of personality, and dose of amytryptaline on the reduction of depression are typical examples for regression analysis.

In *linear regression analysis,* a straight line passing through the dots in dot diagram is to be fitted as shown in Figure 5.3 (for the data described for calculation of r).

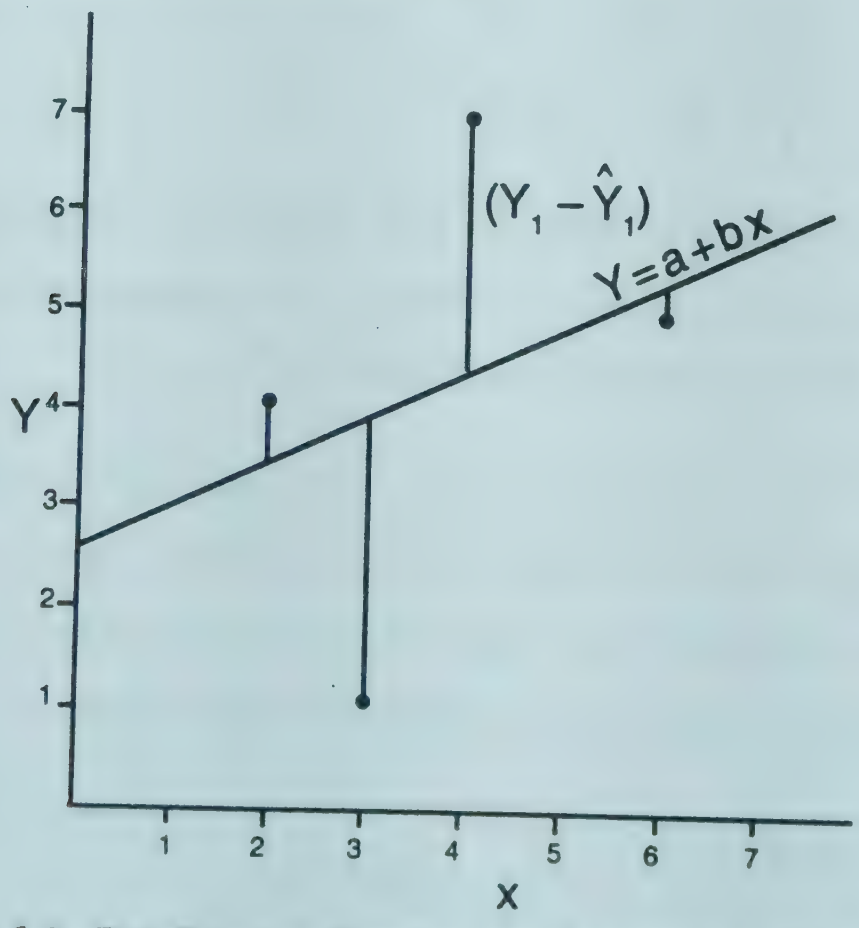


Figure 5.3 Dot Diagram Showing the Regression Line of y on x.

The straight line equation is given by,

$$y = a + bx$$

where 'a' is the intercept and 'b' is the slope of the straight line. The parameters a and b are determined in such a way that the sum of squares of deviations of the observed values from their regression estimates is minimum. This is a typical problem of differential calculus. The normal equations which yield the least square solution are given by,

$$\Sigma\ (y-a-bx) = 0$$

and

$$\Sigma\ x\ (y-a-bx) = 0$$

Solving for a and b gives,

$$a = (\ \bar{y} - b\ \bar{x}\ )$$

and

$$b = \frac{\Sigma\ xy - n\ \bar{x}\ \bar{y}}{(\ \Sigma x^2\ - n\ \bar{x}^2\ )}$$

Hence,

$$\hat{y} = (\ \bar{y}\ - b\bar{x}\ ) + bx$$

$$= \bar{y} + b\ (\ x - \bar{x}\ )$$

Where $\hat{y}$ is the predicted value of the depended value for the given value of independent variable x. This is the regression equation of y on x and b is the regression coefficient of y on x.

Let us consider the data given for the calculation of correlation coefficient as a regression analysis problem. The x may be the number of theory classes attended by a group of five students and y may be the marks scored in bio-statistics test. For this data,

$$b = \frac{(69-5 \times 3 \times 4)}{(65 - 5 \times 3^2)} = \frac{9}{20} = 0.45$$

and $a = 4 - 0.45 \times 3 = 2.65$

Hence $\hat{y} = 2.65 + 0.45\,x$

The observed scores (y) and predicted scores ($\hat{y}$) of the dependent variable are computed as presented below.

Students	x	y	$\hat{y}$	$(y - \hat{y})$
A	4	7	4.45	2.55
B	2	4	3.55	0.45
C	3	1	4.00	-3.00
D	0	3	2.65	0.35
E	6	5	5.35	-0.35
Total	15	20	20.00	0.00

Suppose the researcher is intend to predict the marks scored in bio-statistics of a student who has attended only one theory class. Hence,

$$\hat{y} = 2.65 + 0.45 = 3.10 \cong 3 \text{ marks}$$

CHAPTER 6

RELIABILITY AND VALIDITY OF MEASUREMENTS

In construction of rating scales, the included items must be reliable, valid, discriminative and moderately difficult. The measurement of these properties and the *item analysis* forms an important aspect.

(a) Reliability of Measurements

The reliability is the *consistency* with which a measure assesses a given trait. In psychological, sociological and educational measurements, reliability depends upon the population measured as well as upon the measuring instrument. Logically, the *reliability coefficient* of any set of measurement may be defined as the proportion of their true variance to the total variance. Each single measurement has two components, viz., a true measure and an error attached to it. It is assumed that the error components occur independently and at random and their mean is zero. Thus, we can treat the observed measure as dependent variable and the true measure as independent variable, and they may be subjected to regression analysis. The correlation between the two components is known as the index of reliability.

There are many procedures to estimate the reliability coefficient from empirical data, falling roughly into retest reliability and internal-consistency reliability. The *test-retest method* consists in submitting a group of individuals to a particular test and compiling their respective scores.

After some time, the same test is repeated on the same group of individuals and their scores are noted down. Then the correlation between the two sets of related scores is obtained to measure the reliability of the test. If the test is repeated immediately after the first, then the scores are likely to be improved on account of the memory effect, practice and confidence. If sufficient time is given, then some other factors may come in such as growth in case of children which may increase the scores. This method is generally used to measure the reliability of speed tests where speed is an important criterion. It is appropriate when the test is heterogeneous in the sense that different parts measure different traits.

The *internal-consistency reliability* is usually measured by split-half method or the method of rational equivalence. These methods are appropriate for power test in which all examinees have time to finish. The *split-half method* consists in breaking the original test into two equivalent halves and computing the correlation coefficient (r_{hh}) between the scores in half tests. Then the reliability coefficient (r_{tt}) for the whole test is determined in terms of self correlation of the half tests by using Spearman-Broken prophecy formula,

$$r_{tt} = \frac{2\, r_{hh}}{1 + r_{hh}}$$

In case of power-test, the test items are arranged in increasing order of difficulty and then splitting them into two equivalent halves with odd and even numbered items provide an unique estimate of reliability.

The *method of rational equivalence* or the Kuder-Richardson method is used by teachers and others who want to determine quickly the reliability of short objective classroom type tests. This method stresses the importance of the inter-correlations of the test items and the correlations of the items with the test as a whole. The method is based on the

assumption that all the items of the test are of equal or nearly equal difficulty. In this method, the formula for determining the test reliability is given by,

$$r_u = \frac{k}{(k-1)} \left[\frac{\sigma_t^2 - \sum p_i q_i}{\sigma_t^2} \right]$$

Where k is the number of items in the test, p_i is the proportion of the i^{th} item answered correctly, $q_i = 1 - p_i$, and σ_t is the standard deviation of individual test scores. The expression $\sum p_i q_i$ is the sum of the variances of all items. Deducting this quantity from the total test variance, left with the sum of covariance. It is in these covariance that the source of true variance lies. Let us suppose that the scores of five students on six items (1 for correct answer and 0 for incorrect answer) are given below.

	Items						
Students	1	2	3	4	5	6	Total
1	1	1	0	0	1	1	4
2	0	1	0	0	0	1	2
3	0	1	0	0	1	1	3
4	0	0	0	0	0	0	0
5	1	1	1	1	1	1	6
Total	2	4	1	1	3	4	15
p_i	0.4	0.8	0.2	0.2	0.6	0.8	-
q_i	0.6	0.2	0.8	0.8	0.4	0.2	-
$p_i q_i$	0.24	0.16	0.16	0.16	0.24	0.16	1.12

For this data, $k = 6$, $\sum p_i q_i = 1.12$ and $\sigma_t = 2$.

Hence, $$r_{tt} = \frac{6}{5} \left[\frac{4 - 1.12}{4} \right] = 1.2\ (2.88/4) = 1.2\ (0.72) = 0.864$$

(i) Inter-rater Reliability

Even if a concept such as schizophrenia is validated by certain genetic, biochemical, psychological, course or treatment variables, the inter-rater reliability assessment of such a concept is required to assume communicative value and to provide the foundation for scientific evaluation. The inter-rater reliability of *two raters and dichotomous rating* is the simplest measurement of inter-rater reliability. The data format is as illustrated below.

Rater A	Rater B		Total
	Schizophrenia	Not Schizophrenia	
Schizophrenia	40	20	60
Not Schizophrenia	10	30	40
Total	50	50	100

The inter-rater reliability may be measured by using *kappa coefficient* (k) given by,

$$k = \frac{(p_0 - p_c)}{(1-p_c)}$$

where p_0 is the proportion of the sum of the main diagonal entries, and p_c is the proportion of the sum of the chance number of agreements. The kappa coefficient lies between -1 and 1. In our example,

$$p_0 = \frac{(40+30)}{100} = \frac{70}{100} = 0.70$$

The chance frequency for schizophrenia by both rater A and the rater B is given by (50×60)/100=30.The chance frequency for non-schizophrenia by both the raters is given by (50×40)/100=20. Hence the proportion of the chance frequency is given by,

$$p_c = (30+20) / 100 = 0.50$$

Hence $$k = \frac{(0.70 - 0.50)}{(1 - 0.50)} = 0.40$$

The data for *two-raters and polychotomous rating* may also be dealt with by using the method of kappa coefficient as described above.

The data format for *k-raters and dichotomous ratings* (whether the subject is a schizophrenic or not) is as illustrated below.

	Raters			
Subjects	1	2	3	Total
1	0	0	1	1
2	1	0	0	1
3	1	1	1	3
4	0	0	0	0
5	0	1	0	1

Here the reliability is determined by the method of intraclass correlation coefficient (ICC) given by,

$$ICC = 1 - \frac{NK(TK - \Sigma T_j^2)}{T(NK-T)(K-1)}$$

Where N is the number of subjects, K is the number of raters, T is the total agreed responses and T_j is the number of agreed response to the j^{th} subject. In our example, $N = 5$, $k = 3$, $T = 6$ and $\Sigma T_j^2 = 12$.

$$ICC = 1 - \frac{5 \times 3 (6 \times 3 - 12)}{6 (5 \times 3 - 6)(3 - 1)} = 1 - \frac{5}{6} = 0.167$$

(b) Validity of Measurements

The validity of a test is the accuracy with which it measures what it is supposed or *intended to measure*. The validity of a mental test can never

be estimated as accurately as can the validity of a physical instrument. The validity of a test is determined experimentally by obtaining the correlation coefficient between the scores of n individuals on the given test (x) and on some independent standard test (y) called criterion. A criterion may be an objective or a qualitative measure. A high correlation coefficient between x and y is an evidence of validity provided that the criterion was set up independently and both x and y are reliable. For example, the validity of a typing test may be judged by calculating the correlation between the errors (x) in the matter typed and speed of typing (y). A test which is not quite reliable can hardly be valid since the test which correlates poorly with itself cannot correlate well with the measure of any other variable.

It is difficult to measure validity in practice because of several concepts involved in it. It is said to have *face* validity if it runs through all the scale items. A scale is said to have *content* validity if it covers the full range of the attitude and covers it in a balanced way. It relates to the problem of selecting a representative sample of items from the bank of items. The *predictive* validity is concerned with how well the scale can forecast a future criterion and *concurrent* validity with how well it can describe a present one. The *construct* validity is based on the theory. The *factorial* validity of a given test is defined by its factor loadings and these are given by the correlation of the test with each factor.

The more the heterogeneity of the group, the greater will be the test variability and consequently the reliability coefficient is higher. Increasing the length of a test tends to increase its reliability. This increased reliability is determined by Spearman – Brown prophecy formula,

$$r_{nn} = \frac{n\, r_{11}}{1 + (n-1)\, r_{11}}$$

Where r_{11} is the reliability of the given test, and n is the number of times the length of the test is to be increased or decreased. The prophecy formula may be used to determine the number of times a test should be lengthened or repeated in order to obtain a test with specified reliability. It is given by,

$$n = \frac{r_{nn}\,(1 - r_{11})}{r_{11}\,(1 - r_{nn})}$$

A highly valid test cannot be unreliable since its correlation with a criterion is limited by its own index of reliability. To be valid a test must be reliable.

(c) Discriminate Index, Difficult Index

In addition to the validity and reliability, the items selected for rating scales must be discriminative and moderately difficult. The *discriminate index* denoted by Dis(i) of a test item is the power of the item to discriminate between people of favourable attitude and people of unfavourable attitude. The discriminate index of the i^{th} item is given by,

$$\text{Dis(i)} = \frac{\text{Number of correct answers to the } i^{th} \text{ item in the favourable group} - \text{Number of correct answers to the } i^{th} \text{ item in the unfavourable group}}{\text{Number of cases in the favourable group}}$$

Where the favourable group consists of the first 27% and the unfavourable group consist of the last 27% of the individuals arranged in a descending order of marks scored by them. Thus, the discriminate index lies between –1 and +1. The Dis(i) is 1 means that a person who posses that item would be in the favourable group and a person who fails that item would be in the unfavourable group. The discriminate index is said to be satisfactory if it is more than or equal to 0.30. As an

hypothetical example, the scores of eleven subjects on five items are given in the following table. For example, the discriminate index of the first item is given by, Dis(1)= (3–1)/3 = 2/3 = 0.67

The *difficult index* denoted by Dif (i) of a test item indicates the proportion of candidates wrongly answered the item. Thus, the difficult index of the i^{th} item is given by,

$$\text{Dif (i)} = \frac{\text{Number of wrong answers to the } i^{th} \text{ item}}{\text{Total number of responses to the } i^{th} \text{ item}}$$

For example, the difficult index of the first item in the above cited example is (11– 8)/11 = 0.27. The difficult index lies between zero and one. All the items in the scale must have moderate difficult indices. They are constructed in such a way that 50% of the candidates are expected to answer correctly, since the variance (information) is maximum at p = 0.50, which is pq = 0.25. That is, as p approaches to 0 or 1 the variance decreases toward the vanishing point.

	Items					
Subjects	1	2	3	4	5	Total
1	1	1	1	1	1	5
2	1	1	1	1	1	5
3	1	1	1	0	1	4
4	1	1	1	0	1	4
5	1	0	1	0	1	3
6	1	0	0	1	1	3
7	1	0	1	0	1	3
8	0	0	1	0	1	2
9	0	0	1	0	1	2
10	0	0	1	0	1	2
11	1	0	0	0	1	2
Total	8	4	9	3	11	35
Dis (i)	67	1.00	.33	.67	0.	–
Dif (i)	.27	.64	.18	.73	0.	–

CHAPTER 7

LIFETABLE TECHNIQUES AND TIME SERIES ANALYSIS

The *time variable* is an important component in several studies concerned with mental health care research and service . The life table techniques and the time series analysis are the important statistical methods. The life table method deals with the expectation of occurrences of phenomena according to time. The time series analysis deals with the factors influencing variation of several phenomena according to time and thus aids in forecasting events.

(a) Life Table Techniques

It is emphasized in the previous chapter that the average duration of stay of hospitalized patients is an indicator of the expectation of duration of stay of a patient in the hospital at the time of admission to in-patient service. The expectation of further duration of stay of the patient after already stayed for say ten days, twenty days, fifty days etc., may be obtained by constructing *hospital stay table*. The hospital stay table is based on the techniques of life table. Let us suppose that a group of twenty patients who were admitted in an emergency psychiatric ward on a particular day were followed up in time and noted down how many of them remained at the end of the first day, how many of them remained at the end of the second day and so on. From this basic data denoted by l_x, the other components of the hospital stay table are computed as shown in Table 7.1.

Table 7.1 Basic Structure of Hospital Stay Table

x	l_x	d_x	L_x	T_x	e_x
0	20	1	19.5	68.0	3.4
1	19	1	18.5	48.5	2.6
2	18	7	14.5	30.0	1.7
3	11	5	8.5	15.5	1.4
4	6	3	4.5	7.0	1.2
5	3	2	2.0	2.5	0.8
6	1	1	0.5	0.5	0.5
7	0	–	–	–	–

The *x* is *the stay* in days. It is taken as 0,1,2,... days in the present table. The l_x *is the* number of patients stayed up to the end of the x^{th} day. For example, there were 18 patients stayed up to the end of the second day. The d_x *is the* number of patients who got discharged between the x^{th} day and the $(x+1)^{th}$ day. Thus,

$$d_x = l_x - l_{x+1}$$

For example, there were 7 patients who got discharged between the second day and the third day. The L_x *is the* number of days stayed by the cohort between the x^{th} day and the $(x+1)^{th}$ day. It may be estimated by,

$$L_x = l_x - \tfrac{1}{2}\, d_x$$

For example, a total of 14.5 days have been contributed by the patients who stayed between the second day and the third day. The T_x *is the* number of days stayed by the cohort beyond the x^{th} day. Thus,

$$T_x = L_x + L_{x+1} + \ldots$$

For example, a total of 30 days have been contributed by the patients who have stayed beyond the second day. The e_x *is the* expectation of

further duration of stay of a patient who has already stayed for x days. It is estimated by,

$$e_x = \frac{T_x}{l_x}$$

For example, 1.7 days was the further expected stay of a patient who has already stayed for two days. These expected values have several administrative and clinical utilities.

The life table techniques were originally developed in the field of demography to express the duration of life experienced by a particular group of population during a particular period. The admissions, discharges and patients remaining in the hospital stay tables are *analogous* to the births, deaths and living population in the life tables. In life table terminology, the l_x is the number of persons lived by the cohort up to the end of the x^{th} year. The d_x is the number of deaths in the cohort between the x^{th} year and the $(x+1)^{th}$ year. The L_x is the number of years lived by the cohort between x^{th} year and $(x+1)^{th}$ year. The T_x is the number of years lived by the cohort beyond x years. The e_x is the life expectancy of a person who has already lived for x years. The values of x are 0,1,2, –....so on for complete life table and the values of x may be taken like 0,5,10 ....so on for an abridged life table. The l_x is the basic component for cohort life table. The basic component for current life table is q_x which is the proportion out of those live up to x years died before reaching the $(x+1)^{th}$ year. In this case,

$$d_x = l_x \times q_x$$

The expectation of life at birth in India was 57.9 years (57.7 for males and 58.1 for females) during 1988. The Kerala state had the highest expectancy of 69.5 years, while Madhya Pradesh had the lowest expectancy

of 53.0 years. The life table techniques may be applied to specific population subgroups such as gender, domicile and disorders. Several studies revealed that mentally retarded persons do not live as long on the average as do non-retarded individuals.

(i) Modified Hospital Stay Table

A method of modification in the construction of hospital stay table is available to facilitate for making use of the *entire data* collected during the period. Let us suppose that a group of 20 patients were admitted in an emergency psychiatric patient ward on some day, say Monday. Out of these patients, 19 were remained up to the end of Monday, 18 were remained up to the end of Tuesday and so on till there was only one patient left at the end of Saturday, as shown in Table 7.2a. Similarly, the number of patients who were admitted on Tuesday, Wednesday etc., were followed till the end of Saturday, as shown in the table.

Table 7.2a Number of Patients Admitted and Discharged in a Psychiatric Emergency Ward

Day of admission	Number of admissions	Number of patients stayed up to the end of					
		Mon	Tue	Wed	Thu	Fri	Sat
Monday	20	19	18	11	6	3	1
Tuesday	15	–	14	12	5	5	2
Wednesday	18	–	–	16	12	8	4
Thursday	21	–	–	–	17	15	10
Friday	22	–	–	–	–	20	19
Saturday	16	–	–	–	–	–	14

The data in Table 7.2a is rearranged as shown in Table 7.2b for clarification.

Table 7.2b Number of Patients Admitted and Discharged in a Psychiatric Emergency Ward

Day of admission	Number of admissions	Number of patients stayed up to the end of the day					
		1st	2nd	3rd	4th	5th	6th
Monday	20	19	18	11	6	3	1
Tuesday	15	14	12	5	5	2	–
Wednesday	18	16	12	8	4	–	–
Thursday	21	17	15	10	–	–	–
Friday	22	20	19	–	–	–	–
Saturday	16	14	–	–	–	–	–
Total number at the end of the day	–	100	76	34	15	5	1
Number available for discharge on that day	–	112	86	57	24	11	3
Probability of staying on that day	–	0.89	0.88	0.60	0.63	0.45	0.33

From the Table 7.2b, it may be noted that the number of patients admitted and available for discharge on the first day was 112. Out of these 100 patients stayed up to the end of the first day. Therefore, the probability of staying up to the end of the first day is 100/112 = 0.89. There were 86 (that is,100-14) patients available for discharge on the second day, and there were 76 patients at the end of the second day. Therefore, the probability of staying up to the end of the second day was 76/86 = 0.88, and so on. From this basic component, denoted by p_x, the other components may be computed as already described in the construction of cohort hospital stay table. With $l_0 = 100$, the results are presented in Table 7.2c.

Table 7.2c Hospital Stay Table for Patients in a Psychiatric Emergency Ward

x	p_x	l_x	d_x	L_x	T_x	e_x
0	0.89	100	11	94.5	311.0	3.11
1	0.88	89	11	83.5	216.5	2.43
2	0.60	78	31	62.5	133.0	1.71
3	0.63	47	17	38.5	70.5	1.50
4	0.45	30	17	21.5	32.0	1.07
5	0.33	13	9	8.5	10.5	0.81
6	–	4	4	2.0	2.0	0.50

The hospital stay tables may be constructed for males and females separately or for different patients groups such as wards, diseases and units.

(ii) Clinical Applications

The life table techniques may be applied to follow-up data of specific psychiatric patients maintained with certain *specific drug*. The consultant psychiatrist may ask questions like 'what are the chances that my patient will remain asymptomatic during the next one year, two years, etc'. The applications of life table techniques will provide the answers using data from patients on whom the follow-up information is available.

(b) Time Series Analysis

A time series is a set of statistical observations arranged in *chronological order*. The monthly number of registrations in a mental hospital over a period of ten years, and the daily bed occupancy of the hospital over a period of five years are typical examples for time series data. Any time series is composed of four components, viz., the trend denoted by T, cyclical variation C, seasonal variation S, and irregular fluctuation I. These components could be related in any manner, but the relationship is

usually considered to be additive. That is, some observation Y is composed of the components given by,

$$Y = T + C + S + I$$

The *trend* of the time series indicates whether the series increase or decrease. It may be determined by fitting a least square straight line equation to the data. That is, under usual notations,

$$T = a + bx$$

Where x represents the time variable given serially from 1 to the number of observations. The trend is subtracted from the observations in order to obtain the variation that may be explained by the sum of the cyclical variation, seasonal variation and irregular fluctuation. That is,

$$(Y–T) = C + S + I$$

The *seasonal variation* represents that portion of a series which may be attributed to the time of the year. The number of registrations in a mental hospital during a particular month may depend on the whether conditions, school vacations and agricultural activities during the month. The actual prevalence of mental disorders in the community during the month has considerable effect. Often the arithmetic means of the seasons are calculated from the residues C+S+I to represent the seasonal variation. The seasonal variation is subtracted from C+S+I , in order to get the sum of the cyclical variation and irregular fluctuation. That is,

$$(Y– T– S) = C+I$$

The *cyclical variation* is a phenomenon that is observed more frequently than the duration of the time series. The phenomena may be attributed to the administrative reforms of the hospital, natural calamity in the surrounding community etc. It may be determined by the method of moving average to the residuals C+I. The cyclical variation is subtracted from the residuals C+I to obtain the *irregular fluctuation.*

(i) Monthly Number of Registrations

In order to *demonstrate* the computations of time series components and to use them in forecasting, let us deal with the monthly number of registrations of dissociative disorder patients over a period of five years at a general hospital psychiatric unit. As the number days of the months vary alternatively, the figures were added up for bimonthly as presented in Table 7.3a. A scanning of the figures in the last column of the table concludes an upward trend of the series. The series gradually decrease up to the year 1998 and increase thereafter. A scanning of the figures in the last row of the table indicates a seasonal peak during March-April.

Table 7.3a Bimonthly Number of Registrations of Dissociative Disorder Patients at a General Hospital Psychiatric Unit

Year	Jan–Feb	Mar–Apl	May–Jun	Jul–Aug	Sep – Oct	Nov–Dec	Total
1996	5	7	3	5	3	6	29
1997	1	6	7	4	3	5	26
1998	2	4	6	2	4	2	20
1999	4	7	7	5	5	5	33
2000	2	9	6	10	7	2	36
Total	14	33	29	26	22	20	144

In this time series data, x is taken from 1 (Jan–Feb 1996) to 30 (Nov-Dec, 2000). In this data, $\bar{x} = 15.5$, $\bar{y} = 4.8$, $\Sigma x^2 = 9455$, and $\Sigma xy = 2355$. It may be computed that, $b = 0.055$ and $a = 3.952$. Thus, the trend is given by , $T = 3.952 + 0.055x$. The trend along with the data are presented in Figure 7.1. The trend is subtracted from the observations in order to get the sum of the cyclical variation, seasonal variation and the irregular fluctuation, as shown in Table 7.3b.

Table 7.3b Sum of Cyclical Variation, Seasonal Variation and Irregular Fluctuations of Number of Registrations of Dissociative Disorder Patients

Year	Jan–Feb	Mar–Apr	May–Jun	Jul–Aug	Sep – Oct	Nov–Dec
1996	1.00	2.94	–1.11	0.83	–1.22	1.72
1997	–3.33	1.62	2.56	–0.49	–1.55	0.40
1998	–2.66	–0.71	1.23	–2.82	– 0.88	– 2.93
1999	–0.99	1.96	1.90	–0.15	– 0.21	– 0.26
2000	–3.31	3.63	0.58	4.52	1.47	– 3.59
Mean (seasonal)	-1.86	1.89	1.03	0.38	– 0.48	– 0.93

The seasonal component is subtracted from the data in Table 7.3b in order to obtain the sum of the cyclical variation and the irregular fluctuation, as shown in Table 7.3c.

Table 7.3c Sum of the Cyclical Variation and Irregular Fluctuations of Number of Registrations of Dissociative Disorder Patients

Year	Jan–Feb	Mar–Apl	May–Jun	July–Aug	Sep–Oct	Nov–Dec	Mean (cycle)
1996	2.86	1.05	–2.14	0.45	–0.74	2.65	0.69
1997	–1.47	–0.27	1.53	–0.87	–1.07	1.33	–0.14
1998	–0.80	–2.60	0.20	–3.20	–0.40	-2.00	–1.47
1999	0.87	0.07	0.87	–0.53	0.27	0.67	0.37
2000	–1.45	1.74	-0.45	4.14	1.95	-2.66	0.55

The graphical representation of 12–month moving average curve showing the cyclical variation as well as the seasonal variation is also drawn in Figure 7.1. Suppose the practitioner of the psychiatric unit is intend to forecast the number of registrations of dissociative disorder patients for the period of January–February, 2002. Then

$$\hat{y} = (3.952 + 0.055 \times 37) + (1.01 \times 1) - 1.86 = 5.137 \cong 5 \text{ Patients.}$$

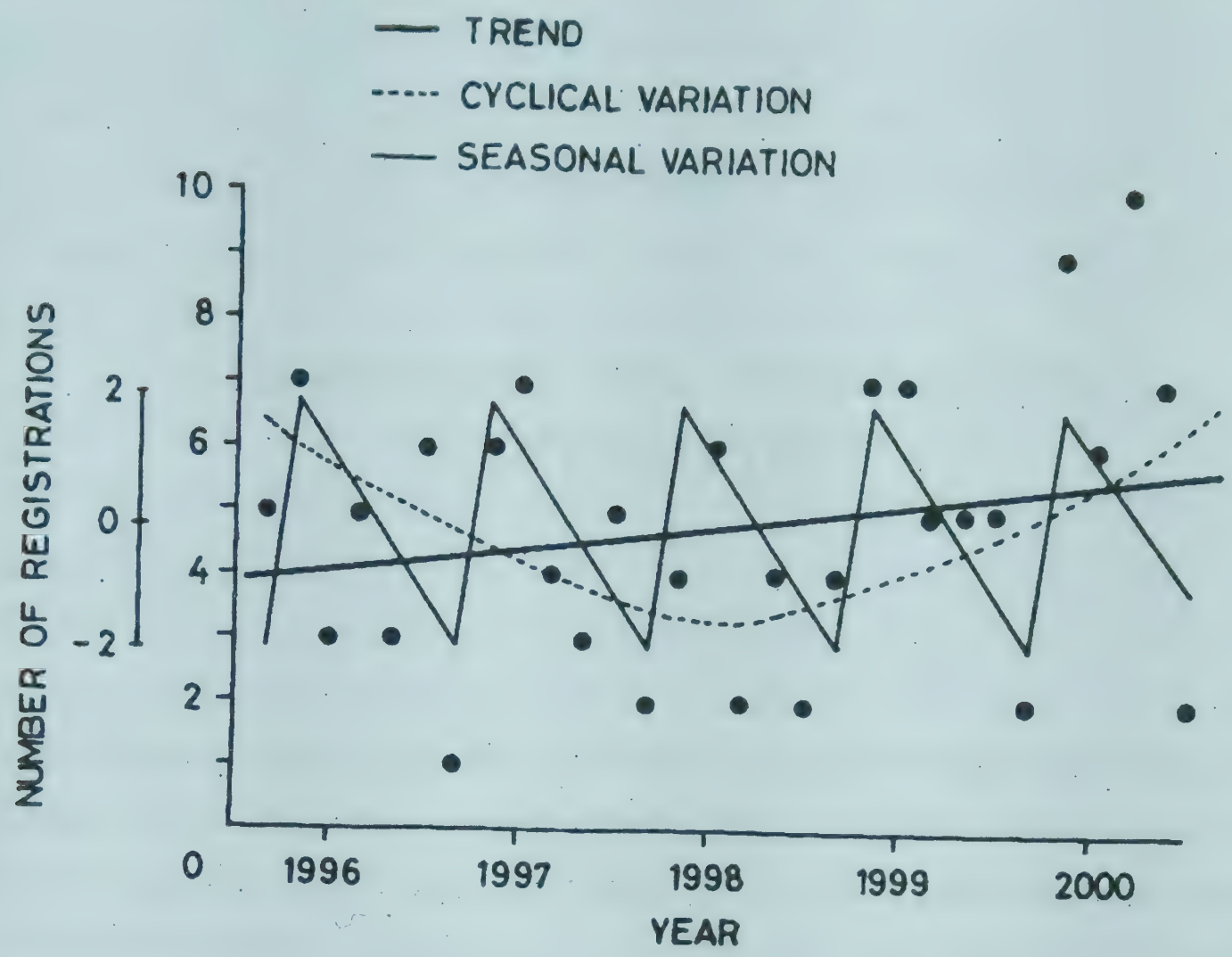


Figure 7.1 : Time Series Analysis Showing the Trend, Cyclical Variation, and Seasonal Variation of Bimonthly Number of Registrations of Dissociative Disorder Patients.

CHAPTER 8

PROBABILITY AND PROBABILITY DISTRIBUTIONS

The *statistical inference* is the science of drawing valid conclusions on the population based on the information contained in the sample (inductive inference). The descriptive statistics forms the material for inferential statistics and the theory of probability is the basis for such inferences. Chevilier DeMoivre has given the idea of defining the term probability (chance) after noticing that certain events associated with gambling occurred in systematic measures. Bernoulli gave the first definition of probability, which was contained in his book published posthumously.

There are mainly two concepts attached to the term probability, viz., the A Priori Probability (mathematical or theoretical probability) and the A Posteriori Probability (statistical or empirical probability). If there are n possible cases associated with an event E, all of them are equally likely and m of them are favourable to E, then the ratio $\frac{m}{n}$ is defined as the *A Priori Probability* of E. That is,

$$\text{Probability (E)} = P(E) = \frac{\text{Number of favourable cases}}{\text{Number of possible cases}}$$

In tossing a coin, there are two possible cases, viz: the occurrence of Head, occurrence of Tail. Let us suppose that the occurrence of Head is

favourable to us. Then, P(Head) = ½ = 0.50 or 50%. This definition is circular, the assumption of equally likelihood itself involves in the concept of probability. Again, there are events outside the field of games of chance for which we cannot make the assumption of equally likelihood, such as the chance that a patient gets cured within one month and the chance that the patient gets cured within two months, etc.

The mental health care research is mainly concerned with the A Posteriori Probability. Let an indefinite number of trials be made with reference to the occurrence of an event E. If the trials are all mutually independent and in the m occasions out of a total of n trials E has been observed, then the *A Posteriori Probability* of E is defined as the limit of $\frac{m}{n}$ as n tends to infinity. That is,

$$P(E) = \underset{n \to \infty}{\text{Limit}} \frac{m}{n}$$

There are 124 schizophrenics out of a total of 500 consecutively registered psychiatric patients at NIMHANS hospital, Bangalore. The probability that a registered patient happens to be schizophrenic is obtained by 124/500=0.248 or 24.8%. The limit that is used here is different from the one that we use in the field of calculus. What we mean by this limit is that we write 1 whenever the event occurs and we write 0 whenever the event is not occurred, and thus we write like 00100001001000.... so on and the probability is calculated. Thus the A Posteriori Probability (frequency probability) is the ratio of the number of times the outcome occurred to the total number of trials.

The probability is the *degree of confidence* in the occurrence of events. The probability is the *measure of occurrence of events*. This measure being a number lies between 0 and 1. The probability P=0 implies that the occurrence of the event is impossible, P=1 implies that the occurrence of the event is certain, and P=0.50 means 50% chance for the event to occur.

(a) Laws of Probability

The addition law of probability is based on the condition that the events are mutually exclusive. Two events are said to be mutually exclusive if the occurrence of one of them prevents the occurrence of the other, that is they do not occur simultaneously. In collecting data on gender of a patient, the recording of male and the recording of female are mutually exclusive. The *addition law* of probability states that, if two events A and B are mutually exclusive, then the probability of the event 'A or B' is given by the sum of their probabilities. That is,

$$P(A \text{ or } B) = P(A) + P(B)$$

For example, it is found that 0.05% of psychiatric registrations at NIMHANS hospital have pre-menstrual tension (PT), and that 0.05% of the registrations have electile dysfunction (ED). Then the probability that a registered patient happens to be either pre-menstrual tension or electile dysfunction is obtained as, $P(PT) + P(ED) = 0.0005 + 0.0005 = 0.001$ or 0.1%.

The multiplication law of probability is based on the condition that the events are independent. Two or more events are said to be independent if the outcome of the first trial does not effect the outcome of the second trial and so on. The *multiplication law* of probability states that, if two events A and B are mutually independent, then the probability of the event 'A and B' is given by the product of their probabilities. That is,

$$P(A \text{ and } B) = P(A) \times P(B)$$

Let us consider two consecutively registered psychiatric patients at NIMHANS hospital. The probability that both of them happen to be schizophrenics is obtained as,

$$P(\text{schi \& schi}) = P(\text{schi}) \times P(\text{schi}) = 0.248 \times 0.248 = 0.0615 \text{ or } 6.15\%$$

If A is an event, not-A is called its complementary event and it is denoted by A′. Its probability called *complementary probability* is given by,

$$P(A') = 1 - P(A)$$

The probability that a registered patient at NIMHANS hospital happens to be non-schizophrenic is obtained by, P(non-schizophrenic) = 1– P(schizophrenic)= 1- 0.248 = 0.752 or 75.2%.

If the event A can occur only when it is known that B has occurred, then the occurrence of A is conditional to the occurrence of B. The *conditional probability* of A given that B has already occurred is given by,

$$P(A/B) = \frac{P(AB)}{P(B)}$$

Where P(AB) is the probability of occurrence of the event. It is found that 0.4% of psychiatric registrations at NIMHANS hospital have puerpheral psychoses, and that 40% of the registered patients are females. Then, the probability that a registered patient has puerpheral psychoses given that the patient is a female is obtained by, P(puerpheral psychoses/ female) = 0.004/0.40 = 0.01 or 1%.

(b) Bayes' Theorem

Generally, the psychiatrist knows the probability of occurrence of a particular symptom (s) in a patient with a particular psychiatric disorder (d). However, it is *important* to know the probability of occurrence of a particular psychiatric disorder in a patient who has a particular symptom. A theorem attributed to Thomas Bayes may be used to provide the latter probability from the former probability. This theorem may be *derived* from the following expressions,

$$P(d/s) = \frac{P(ds)}{P(s)}$$

$$= \frac{P(d) \times P(s/d)}{P(ds)+P(d's)}$$

$$= \frac{P(d) \times P(s/d)}{[P(d) \times P(s/d)] + [P(d') \times P(s/d')]}$$

It is found that 28% of children registered at the child guidance clinic of NIMHANS hospital have conduct disorder. Consider the two events that a child has conduct disorder (d) and that the child has non-conduct disorder (d′). Since d and d′ are complementary, P(d) = 0.28 and P(d′) = 0.72. It is further known that 63% of the conduct disorder children and 17% of the non-conduct disorder children of the clinic have truancy at home/school. In notation, P(s/d)= 0.63 and P(s/d′)=0.17. Then, the probability that a registered child is conduct disorder knowing that the child has truancy is obtained as,

$$P(\text{conduct disorder/truancy}) = \frac{0.28 \times 0.63}{(0.28 \times 0.63)+(0.72 \times 0.17)} = 0.59 \text{ or } 59\%$$

(c) Probability Distributions

The statistical inferences are drawn by considering sampling distributions and calculating probabilities. The *sampling distributions* differ according to the type of the characteristic studied, the nature of the population and the size of the sample. For each type of situation, a sampling distribution may be formed by using a mathematical model called *theoretical distribution*. In several situations, the observed sampling distributions are very close approximation of the theoretical distributions. The mathematical models have been developed and, the required probabilities were calculated and made available in tables for certain types of distributions as given in Appendix V.

(i) Binomial Distribution

Some times, the mental health workers are interested to know the *proportion* of individuals in a population posses a particular character such as the number of psychiatric patients in a family of five members. An estimate of this proportion is calculated based on a suitably drawn sample and the corresponding sampling distribution. In this type of problem, the sampling distribution is given by a theoretical frequency distribution known as binomial distribution. It is given by,

$$P(x) = {}^{n}c_{x}\, p^{x} q^{n-x}$$

Where n is the sample size, x is the number of individuals posses the event, p is the probability of occurrence of the event and q is the probability of non-occurrence of the event. The ${}^{n}c_{x}$ denotes the number of combinations in a set of n object taken x at a time. The combinations do not take into account of the order of arrangements of the objects. The number of combinations is given by,

$${}^{n}c_{x} = \frac{n!}{(n-x)!\, x!}$$

Where x! = 1×2× ...× x. The prevalence of mental/behavioural disorders in India is estimated to be 5.82% (P=0.0582). A sample of two persons in a family can be any one of the three types: having no psychiatric patient, one psychiatric patient, or two psychiatric patients. The probabilities of these compositions are obtained by using the addition and multiplication laws of probability as given in the following table.

Composition of the sample	Ways of occurrence	Probability of occurrence	Probability of getting such a composition
No psychiatric patient	N N	0.9418 × 0.9418	0.8870
One psychiatric patient	N P	0.9418 × 0.0582 }	0.1096
	P N	0.0582 × 0.9418 }	
All psychiatric patients	P P	0.0582 × 0.0582	0.0034
Total	–	–	1.0000

P – psychiatric patient, N – normal

These probabilities may also be obtained by using binomial distribution. For example, $P(1) = {}^{2}C_{1}(0.0582)^{1}(0.9418)^{1} = 2\times0.058\ 2\times0.9418=0.1096$.

(ii) Poisson Distribution

There are situations in which the number of times an *event occurs* can be counted but the number of times the event did not occur cannot be counted. For example, the number of follow-ups made by a psychiatric patient may be counted but not the number of follow-ups not made. The probabilities of observing no follow-up, one follow-up, two follow-ups and so on in a given sample of such patients can theoretically be found out by the use of poisson distribution. It is given by,

$$P(x) = e^{-m}\frac{m^{x}}{x!}$$

The x is the number of times the event occurs and m is the mean of the distribution. The observed frequency distribution of the number of follow-ups of 100 patients registered at a general hospital psychiatric unit is given in the following table. The mean of the distribution is 0.71. The theoretical probabilities based on poisson distribution and the expected frequencies are worked out and presented in the following table. For example,

$$P(3) = e^{-0.71} \times \frac{(0.71)^3}{3!} = 0.4916 \times \frac{0.3579}{6} = 0.029$$

Number of follow-ups	Observed frequencies	Theoretical probabilities	Expected frequencies
0	51	0.492	49
1	34	0.349	35
2	10	0.124	12
3	3	0.029	3
4	2	0.006	1
Total	100	1.000	100

(iii) Normal Distribution

The binomial and poisson distributions deal with the occurrence of discrete events such as the number of psychiatric patients, the number of follow-ups etc. In many situations, the characteristics to be studied is *continuously* measured such as age, intelligence quotient, body temperature etc. For such variables, many populations and also their sampling distributions are very close to a pattern of frequency distribution known as normal distribution. The mathematical function that generates the probabilities is given by (with usual notations),

$$f(x) = \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left[\frac{-(x-\mu)^2}{2\sigma^2}\right]$$

An *ideal normal distribution* curve is symmetrical. About 68.3% of the observations lie within one standard deviation from either side of the mean, about 95.5% of the observations lie within two standard deviations from either side of the mean, and about 99.7% of the observations lie within three standard deviations from either side of the mean. Thus, the values that differ from the mean by more than three times the standard deviation are very rare.

If x is a normal variate, then the *standard normal variate* or relative normal variate (Z) is defined as,

$$Z = \frac{x - \bar{x}}{SD}$$

The mean of the standard normal values in a data set is zero and the standard deviation is one. Thus, about 95.5% of the standard normal values of the observations lie between –2 and 2. About 95% of the standard normal values of observations lie between –1.96 and 1.96. That is, about 2.5% (P=0.025) of the standard normal values of observations lie above 1.96, and about 2.5% (P=0.025) of the values lie below –1.96, as shown in Figure 8.1.

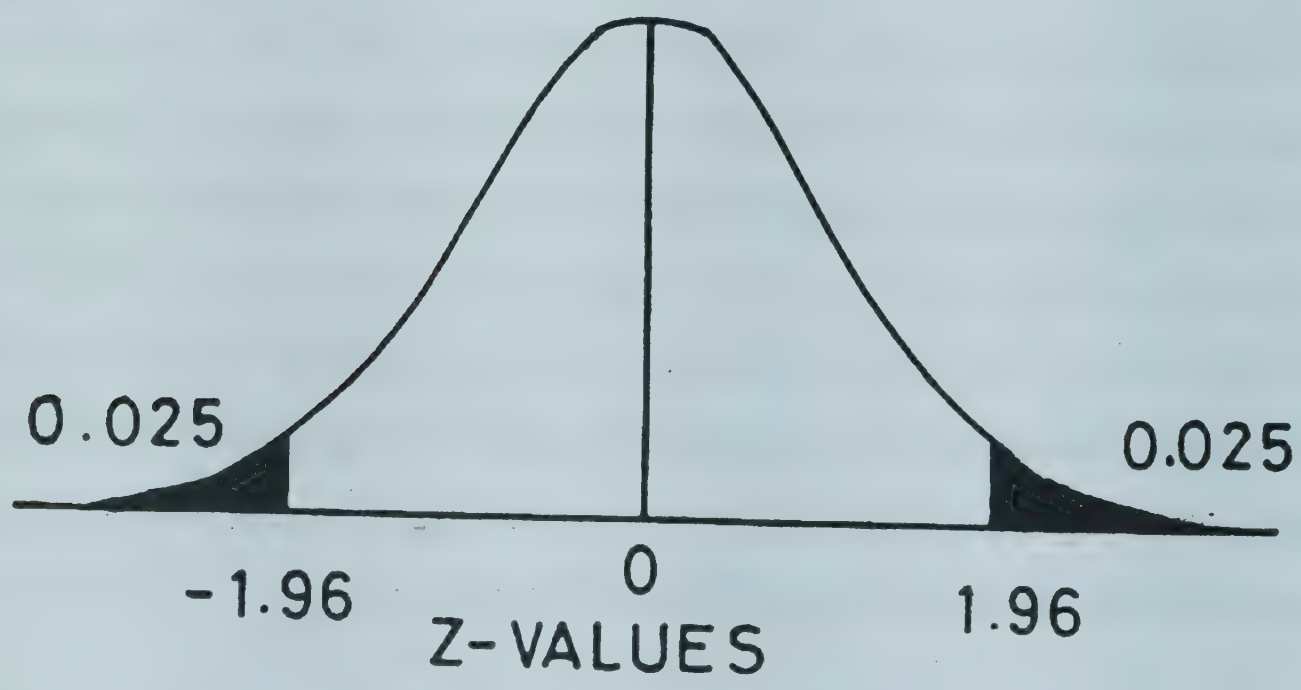


Figure 8.1 Probability of Standard Normal Value of an Observation falling below –1.96 and the Probability of the Value falling above 1.96.

About 99% of the standard normal values of the observations lie between –2.58 and 2.58. The proportions under the entire standard normal curve which lie between zero and a positive value of z are presented in table as given in Appendix V. The proportions of areas between zero and negative values for Z are obtained by symmetry. It is found that the mean age and the standard deviation of a sample of schizophrenics at the time of

registrations in NIMHANS hospital are 33.3 years and 12 years respectively. Then the probability that a registered schizophrenic patient happens to be aged 50 years and above may be worked out. Here, $Z = (50 - 33.3)/12 = 1.392$. Referring to the table value of z (Appendix V), the proportion of area between 0 and 1.392 of z is given to be 0.418. Hence, the proportion of area above 1.392 is (0.500–0.418) = 0.082. Thus, the probability that a registered schizophrenic patient happens to be aged 50 years and above is 0.082 or 8.2%

Many variables encountered in mental health care research have *approximate* normal distribution. If the distribution is approximately normal, no essential details are lost by considering it to be normal. The mean and standard deviation describe the normal distribution completely. The sampling distributions of the statistics calculated from samples taken from the parent normal distribution are known. If we take a series of samples of fixed size say n from the population, then the sample means cluster very closely around the population mean. The standard deviation (SD) of statistics are known as *standard error* (SE). The standard error of mean is given by,

$$SE(\bar{x}) = \frac{SD}{\sqrt{n}}$$

Thus, the sampling error decreases with increase in the sample size. The *central limit theorem* states that, the statistics of all distributions follow normal distribution provided the sample size is sufficiently large.

Several probability distributions have been developed from the parent normal distribution, such as chi-square distribution, student's t distribution and F-distribution. The *chi-square* is the sum of squares of n independent standard normal variates and it is denoted as χ^2. Given a fixed total in a contingency table, the number of independent cells or

cells that can be varied freely will be the number of *degrees of freedom* (df). The *t-distribution* is the ratio of a standard normal variate to the square root of the chi-square divided by their respective degrees of freedom. Here, the size of the sample minus the number of parameters estimated is called the degrees of freedom. Thus it is the number of independent members in the sample. The *F-distribution* is the ratio of two independent chi-squares divided by their respective degrees of freedom. Thus this distribution has two independent degrees of freedom one attached to the numerator and the other attached to the denominator. These probability distributions are presented in tables and their utility are essential in science of statistics for mental health care research.

CHAPTER 9

SAMPLING THEORY AND METHODS

In case the population size is very small, it is convenient to obtain the information by collecting the data on all the units in the population. Most frequently, *time and money may be saved* by measuring only a sample of units. The sampling approach has wide applications in monitoring and evaluation of mental health programs. It is not possible to include all people of India to find the prevalence of mental and behavioural disorders in the country. In sample surveys, greater attention may be paid to each unit selected, completeness and accuracy may be achieved by more persistent efforts and better organization. The sampling is unavoidable in case the population is infinite or hypothetical.

Generally, we are interested in *representative sample* to draw valid conclusions on the population including the estimations of parameters. Only the *random sampling* will assure such representation. In random sampling, every unit in the population has a known probability (chance) of being included in the sample. The sample is drawn by some method of random selection consistent with these probabilities and we take into account of these probabilities while estimating the parameters. The random sampling has several advantages. In random sampling, it is possible to study the bias and error attached to the estimates from different sampling techniques. In this way, much has been reported

about the scope, advantages and limitations of each sampling technique so as to choose the one that suits our sampling job reasonably well. The *non-random sampling* techniques such as convenient sampling, judgment sampling and quota sampling do not assure fair representation of the population, but they must be based on certain specific purposes.

There are *sampling errors* due to sample enumeration instead of the census enumeration. The sampling errors are the errors of estimate arise solely from the sampling variation that is present whenever a sample of N units are measured instead of the complete population of N units. The sampling errors may be reduced by selecting suitable random sampling method, fixing optimum sample size, and eliminating sampling bias.

(a) Random Sampling Methods

Several random sampling methods are available *depending upon* the objectives of the study, the size and nature of the population, required precision of the estimate, sanctioned budget and the availability of the sampling frame. A sampling frame is a list of all the units of the population listed alphabetically or chronologically and numbered serially. A sampling frame must be exhaustive and up-to-date. The simple random sampling, systematic sampling, stratified sampling and cluster sampling are the commonly used random sampling methods in mental health care research field. The ***simple random sampling*** is the simplest of all the random sampling methods. In simple random sampling, each unit in the population has the same chance of being included in the sample at the first draw or at each subsequent draws. The sampling may be carried out by using random number tables, or by lottery system of drawing. *A list of random numbers* is given in Appendix IV. This sampling method is appropriate when the population size is small, sampling units are homogeneous and the frame is readily available, as in the case of clinical trials. Here, the sampling without replacement provides more

efficient estimates than that with replacement. An unbiased estimate of the population mean is given by,

$$\bar{y}_{srs} = \frac{1}{n} \Sigma y_i$$

An estimate of the variance of the estimator is beyond the scope of this book.

The *systematic sampling* is operationally more convenient than simple random sampling and at the same time ensures equal probability to each unit in the population of being inclusion in the sample. Let k = N/n. The k is taken as the nearest integer to N/n. Select a random number from 1 to k, say r. Then, the r^{th}, $(r+k)^{th}$,... $[r+(n-1)k]^{th}$ units are selected for the sample. The k is known as sampling interval which is the reciprocal of the sampling fraction and the random number r is known as random start. This plan is more appropriate when the population size is large, the sampling units are heterogeneous and the sampling frame may not be available in advance, as in the case of surveying a sample of patients attending a psychiatric clinic. This sampling procedure is not valid when there is a periodicity of a particular event under study in the population. An unbiased estimate of the population mean is given by

$$\bar{y}_{sys} = \frac{1}{n} \Sigma y_i$$

The **stratified sampling** may be preferred to increases the precision of the estimate by reducing the heterogeneity of the population. In stratified sampling, the population of N units is divided into L sub-populations of N_1, N_2,...N_L units which are internally homogeneous. These sub-populations are not overlapping and together they comprise the whole population. The classes into which the population is divided are called

strata. Then each stratum will be sub-sampled and a definite number of units are taken from each stratum as in the case of simple random sampling. For example, in estimating the average bed occupancy in psychiatric units of general hospitals attached to medical colleges in India, a samples of medical colleges have to be drawn from different ownerships such as government, government autonomous, private, municipalities, universities etc. The partial samples of the different strata may be used to estimate the means of several strata from which the overall estimate of the population mean may be obtained. An unbiased estimate of the population mean is given by,

$$\bar{y}_{st} = \Sigma W_h \bar{y}_h, \quad \text{where } \bar{y}_h = \frac{1}{n_h} \Sigma y_{hi}, \text{ and } W_h = \frac{N_h}{N}$$

The n_h is the sample size for the h^{st} stratum. If the estimate in each stratum is an unbiased estimate of the stratum mean, then the overall estimate will be an unbiased estimate of the population mean. Various allocation models are available in fixing the sample size for each stratum. Under proportional allocation, n_h are fixed in such a way that,

$$n_h = \frac{n}{N} N_h$$

The sample in this case is more representative than that from the simple random sampling. The success of the stratified sampling depends on some prior knowledge of the population to be stratified.

The ***cluster sampling*** consists in forming suitable clusters of units and surveying all the units in a sample of selected clusters. By cluster sampling, it is usually meant by sampling of clusters of units formed by grouping neighbouring units or units which can be surveyed together.

For example, surveying a random sample of schools in a particular city to determine scholastic backwardness of first grade students in the city. This sampling scheme may decrease the sampling efficiency though operationally more convenient and less costly than simple random sampling. The units within the clusters should be as heterogeneous as possible. If the intracluster correlation coefficient is negative as some times happens, then cluster random sampling is more precise than simple random sampling. The intracluster correlation coefficient is defined as the covariance between the elements of the same cluster. Let us deal with the case in which all the clusters are of equal size. Suppose a finite population of NM units is divided into N clusters of size M in each cluster. A sample of n clusters are selected from N clusters by using simple random sampling without replacement. An unbiased estimate of the population mean is given by,

$$\bar{y}_c = \frac{1}{n} \Sigma \bar{y}_i \qquad \text{Where } \bar{y}_i = \frac{1}{M} \Sigma y_{ij}$$

Area sampling is close to cluster sampling which is more suitable for mental morbidity surveys. Under area sampling we first divide the total area into a number of smaller non-overlapping areas called geographical clusters, then a number of these clusters are randomly selected, and all units in the selected clusters are included in the sample.

(b) Sample Size

The sampling variance decreases with increase in the sample size, but at the same time the cost of the survey also increases. Hence, in practice, the *optimum* sample size which decreases the sampling error for a reasonable cost has to be determined. The *general relationship* between sample size and sampling error is as shown in Figure 9.1

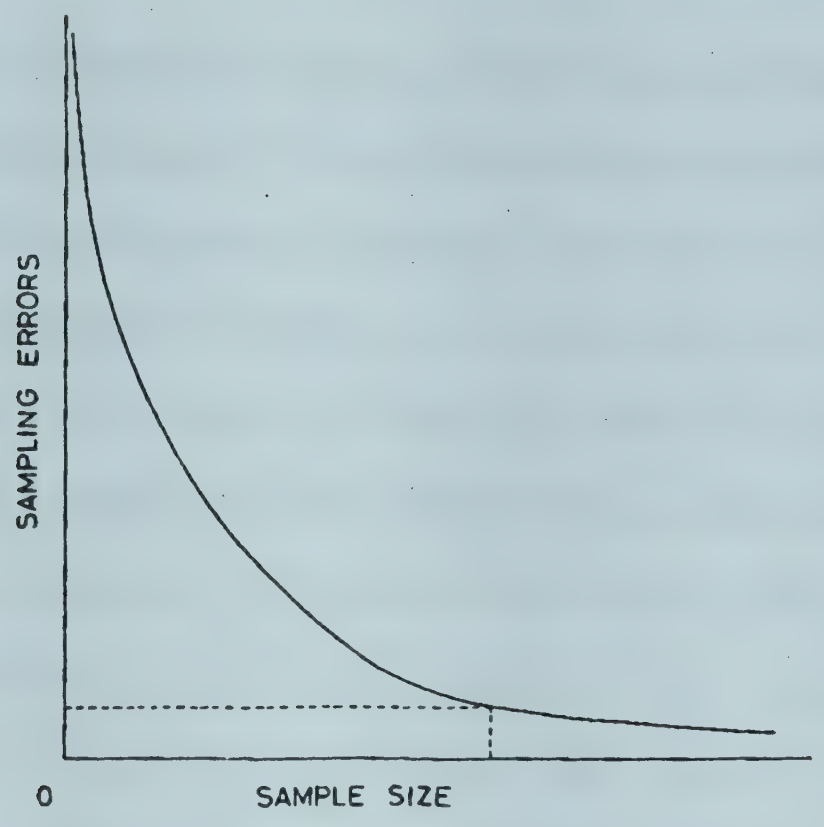


Figure 9.1 General Relationship Between Sample Size and Sampling Error.

In case of estimating *qualitative parameter*, the optimum sample size is given by,

$$n = \frac{Z^2 pq}{L^2}$$

The L is the allowable error due to sampling defects that can be tolerated. It may be fixed at 20% or 10% of the positive character of the qualitative variable. With 5% risk that the true value will exceed allowable error, the Z may be fixed at 1.96. The p is the proportion of the positive character and q = 1 – p. It is established that the prevalence rate of mental and behavioural disorders in India is 58.2 per thousand (5.82% or p = 0.0582) of the population. The sample size required to carry out a mental morbidity study to determine the prevalence rate of a geographical area with 10% allowable error is given by,

$$n = \frac{1.96^2 (0.0582 \times 0.9418)}{(0.0582 \times 0.1)^2} = 6217$$

In case of estimating *quantitative parameter,* the optimum sample size is determine by,

$$n = \frac{Z^2 S^2}{L^2}$$

The S^2 is variance or the estimate of the variance of the study variable. This variance may be obtained from the literature of the subject or by carrying out a pilot study on a small number of units. The case history records of 124 schizophrenia patients at NIMHANS hospital revealed that the mean age of onset of this disorder is 31.0 years with standard deviation of 11.78 years. The age at onset of the illness was obtained by subtracting the duration of illness from the age of the patient at the time of registration. With 10% allowable error, the sample size required to estimate the age of onset of schizophrenia either from the hospital records or from a community survey is determined as,

$$n = \frac{1.96^2 \times 11.78^2}{(31 \times 0.1)^2} = 55$$

(c) Sampling Bias

The *sources* of sampling errors include several types of sampling bias. Some times, it is difficult to demarcate the sample units from the population and the wrong demarcation introduces sufficient bias in the estimate. In case the investigator cannot measure a particular unit selected, he may measure the succeeding or preceding unit in the list, imposing considerable bias in the estimate. Only in sample studies, the problem of estimation arises and hence there may be errors of estimation.

CHAPTER 10

ESTIMATION OF PARAMETERS

The subject matter of *statistical inference* consists of estimation of parameters and tests of significance of hypotheses. An estimator is a function of sample values whereas an estimate is a numerical value of the estimator. An estimate may be a single statistic such as the mean which is called point estimate or a range with attached probabilities called interval estimate or confidence interval.

A good estimator must be consistent, efficient and sufficient. A *consistent* estimator gets increasingly better as the sample size increases. Thus, we expect to get a better estimate of percentage of organic psychotics of registered patients at a psychiatric clinic if we observe 500 patients rather than if we observe only 100 patients. The organic psychotics were 1% in a sample of 100 psychiatric registrations at NIMHANS hospital, they were 3.5% in a sample of 200 registrations, they were 3% in a sample of 300 registrations, they were 2.75% in a sample of 400 registrations and they were 2.8% in a sample of 500 registrations. Both the sample mean and the median are consistent estimators for their respective parameters. An *efficient* estimator is one for which the standard error is small. This is not an absolute criteria since we have not said how

much small the standard error should be. But it allows us to compare two estimators in some circumstances. In many situations, the sample mean tends to be more efficient estimator than the sample median, although there are situations for which the reverse is also true. A *sufficient* estimator makes use of all the potentially useful information in the data. Thus, the sample mean is a sufficient estimator but not the sample median. For example, the mean bed occupancy of psychiatric in-patients of NIMHANS hospital for the year 2000 was 469 patients, which is based on the in-patient census data of all the 366 days of the year. The number of in-patients as on the first day of the year (465 patients), last day of the year (421 patients) or the middle day of the year (474 patients) are not sufficient estimators.

The unbiasedness and resistancy are also important criteria to evaluate an estimator. The expected value of an *unbiased* estimator is equal to the parameter value. The expected value of a random variate is given by,

$$E(x) = \Sigma x_i p_i$$

Where x_i are the sample points and p_i are their attached probabilities. Thus the expected value when a die is rolled is given by, E (die) = $\frac{1}{6}(1+2+3+4+5+6) = 3.5$. The sample mean is an unbiased estimator of the population mean. The mean duration of stay of 23 days of psychiatric patients admitted during the year 2000 at NIMHANS hospital is the expected duration of stay of a psychiatric patient at the time of admission for in-patient treatment in the hospital. An unbiased estimator for population standard deviation is given by,

$$s = \sqrt{\frac{1}{(n-1)} \Sigma (x - \bar{x})^2}$$

A *resistant estimator* is not affected too much by the presence of extreme values or the outliers in the data. Thus the sample median is a resistant estimator but not the sample mean. The median duration of stay of psychiatric patients discharged at NIMHANS hospital during the year 2000 was 15 days which is a resistant estimate. The mean duration of stay of 27 days is not a resistant estimate since it is unduly effected by load of chronic patients. In mental health care research, we should pay particular attention to resistant estimators since gross errors or outliers occur frequently in practice. In some situations, we may prefer a biased estimator when it satisfies all other criteria except being biased.

(a) Evaluation of Screening Tests

In recent years, several screening or diagnostic tests have been established to detect psychiatric cases in the community. Criteria such as the sensitivity and specificity of the test are also established to evaluate these tests. The sensitivity is the ability of the screening test to detect positive cases. That is, the sensitivity is the probability that the screening test indicates a diseased case as positive. The specificity is the ability of the screening test to detect negative cases.

Screening test	Disease	Non-disease	Total
Positive	True Positive (TP)	False Positive (FP)	TP+FP
Negative	False Negative (FN)	True Negative (TN)	FN+TN
Total	TP + FN	FP + TN	n

Thus,

$$\text{Sensitivity} = TP/(TP + FN)$$

$$\text{Specificity} = TN/(FP + TN)$$

(b) Dealing with Sensitive Questions

The estimation of parameters in case of sensitive questions are unduly affected by the presence of high rate of response errors in the data. This requires an indirect way of obtaining data for a reliable estimate of

parameters. The following procedure explain a method of estimating the percentage of pre-marital sex among married women at a psychiatric clinic population.

A. Select a random number from 1 to 10. If it is greater than 3, then answer to question B. Otherwise answer to question C

B. Have you had pre-marital Sex ?

C . Toss a coin. Is it head ?

Let us suppose that there are 25 'yes' responses in a total of 100 trials. Then,

$$\frac{25}{100} = P(\text{B and yes}) + P(\text{C and Yes})$$

$$= \left(\frac{7}{10} \times P\right) + \left(\frac{3}{10} \times \frac{1}{2}\right), \text{ where P is the probability of pre-marital sex.}$$

$$0.7\,P = 0.25 - 0.15 = 0.10$$

$$P = 0.10/0.70 = 0.143 \text{ or } 14.3\%$$

(c) Interval Estimates

The estimation of confidence intervals are important since the investigator can never estimate the exact values of the parameters with certainty. The procedure is to obtain a point estimate and then set up certain limits on both sides of the estimate on the basis of the sampling distribution of the statistics used. Thus, the confidence interval helps in locating the parameter value.

The *sample mean* is the best estimator for population mean. When the sample size is large (greater than 30), the 95% confidence interval for population mean is given by,

$$\bar{x} \pm z_{0.05}\, SE(\bar{x})$$

$$\bar{x} \pm 1.96\,(S/\sqrt{n})$$

Where $z_{0.05}$ is the value in the normal distribution corresponding to P = 0.05. The case history records of 124 schizophrenic patients at NIMHANS hospital revealed that the mean age of onset of this disorder is 31.0 years with a standard deviation of 11.8 years. The 95% confidence interval is obtained as,

$$31.0 \pm 1.96\ (11.8 / \sqrt{124}) \text{ or } 31.0 \pm 2.1 \quad \text{or } 28.9 \text{ to } 33.1 \text{ years}$$

That is, the 95% confidence interval for the mean age of onset of schizophrenic disorders is 28.9 to 33.1 years. When the *sample size is small*, the 95% confidence interval for population mean is given by,

$$x \pm t_{0.005}\ SE\ (\bar{x}) \quad \text{with } df = (n-1)$$

$$x \pm t_{0.005}\ (s/\sqrt{n})$$

Where $t_{0.05}$ is the value of the t-distribution corresponding to P = 0.05 with (n-1) degrees of freedom. The case history records of 14 organic psychotics at NIMHANS hospital revealed that the mean age of onset of this disorder is 39.1 years with an unbiased standard deviation of 14.2 years. The 95% confidence interval is obtained as,

$$39.1 \pm 2.16(\ 14.2 / \sqrt{14}\ ) \text{ or } 39.1 \pm 8.2 \quad \text{or} \quad 30.9 \text{ to } 47.3 \text{ years}$$

That is, the 95% confidence interval for the mean age of onset of organic psychoses is 30.9 to 47.3 years.

The *sample proportion* (p) is the best estimator of the population proportion (P). When the sample size is large, the 95% confidence interval for population proportion is given by,

$$p \pm z_{0.05}\ SE\ (p)$$

$$p \pm 1.96\ \sqrt{pq/n}$$

Based on a sample size of 33572, the prevalence rate of mental and behavioural disorders in India is estimated to be 58.2 per thousand population. The 95% confidence interval is calculated to be

$$0.0582 \pm 1.96 \sqrt{(0.0582 \times 0.9418)/33572} \quad \text{or} \quad 0.0582 \pm 0.0025 \quad \text{or} \quad 0.0557 \text{ to } 0.0607$$

That is, the 95% confidence interval for the prevalence of mental and behavioural disorders in India is 55.7 to 60.7 per thousand population

The sample *correlation coefficient* (r) is the best estimator for population correlation coefficient (ρ). The 95% confidence interval for the population correlation coefficient is given by,

$$r \pm t_{0.05} \, SE\,(r)$$

$$r \pm t_{0.05} \sqrt{(1-r^2)/(n-2)} \qquad \text{with } df = (n-2)$$

The correlation coefficient between the number of years of education and the economic status as measured by the monthly income of 35 psychiatric patients registered at NIMHANS hospital is calculated to be 0.658. The 95% confidence interval for this coefficient is obtained as,

$$0.658 \pm 2.04 \sqrt{(1-0.658^2)/33} \quad \text{or} \quad 0.658 \pm 0.267 \quad \text{or} \quad 0.391 \text{ to } 0.925$$

That is, the 95% confidence interval for the correlation coefficient between the number of years of education and the monthly income is 0.391 to 0.925.

CHAPTER 11

TESTS OF SIGNIFICANCE

A researcher involves in making hypotheses about populations and checking the plausibility of the hypotheses using sample data. The sampling variation is considered in this process. The assumption of no association between factors in a population or no difference between populations is known as *null hypothesis* denoted by H_0. When the null hypothesis is rejected, another hypothesis known as *alternate hypothesis* has to be accepted. The alternate hypothesis is denoted by H_1. The statistic is *significant* means that the difference between the observed score from the sample and the score based on the hypothesis is not due to chance, defying the null hypothesis. On the other hand, the statistic is not significant means that the difference is obtained by chance. In the process of tests of significance of hypotheses, there are chances of committing two types of errors known as the *type I error* and the *type II error*. The type I error is rejecting the null hypothesis when it is actually true. The type II error is accepting the null hypothesis when it is actually false.

Action taken	H_0 is actually	
	True	False
Reject	Type I error	Correct decision
Accept	Correct decision	Type II error

The probability of committing type I error is known as *level of significance.* It is usually fixed at 5% (P=0.05%) or 1% (P=0.01). The *power of the test* is given by,

Power of the test = 1- P (type II error)

Thus the power of the test is the probability of rejecting the false null hypothesis. Experience suggests that in practice the power of the test is set at a maximum of 80%.

The statistical tests of significance which make certain *assumptions* about the populations are known as parametric tests. Such assumptions are: (1) the observations are drawn independently and randomly, (2) the variables involved are measured in at least interval scale, (3) the observations are drawn from normally distributed populations, and (4) the populations have the same variance. Several tests of significance are *robust* in the sense that the probabilities associated with the type I error and the type II error are not grossly affected by the violations of these assumptions. Most of the tests based on the z-distribution, t-distribution and F-distribution are robust. Some times, the original observations may be *transformed* to produce a new set of values so as to facilitate appropriate statistical techniques to the transformed data. The major applications of the transformations are the normalization of distributions, equalization of variances and linearization of relationships.

There are *four steps* involved in the procedures of the tests of significance of hypotheses. They are : (1) stating the null and the alternate hypotheses after analyzing the problem, (2) calculating the standard error of the statistics used, (3) calculating the critical ratio and determining its distribution, and (4) comparing the observed value of the critical ratio with that of the table value at 5% and 1% level of significance and draw valid statistical inference. There are various types of problems for which the tests of significance are used for drawing conclusions. Different types of problems need different tests but the basis and the steps involved in the procedures are the same.

In tests of significance, when we want to determine whether the mean age of schizophrenia is different from that of affective disorders, and does not specify higher or lower, the P-value includes both sides of extreme results, and the test is called *two-tailed test.* In case of 5% level of significance, the P-value will be 2.5% at each end. So the result is compared with table value at the probability level of 0.05. The difference is being tested for significance but the direction is not specified. When we want to test whether the mean age of schizophrenia is less than that of affective disorders, the result will lie at one end of tail of the distribution, and the test is called *one-tailed test.* We specify the direction on plus or minus side. So when the result is compared with table value at P = 0.05,the probability of higher or lower results occurring by chance will be only 0.025.

(a) Tests of Significance on Means

The researcher often involves in making certain hypotheses concerned with a population mean (μ), two populations means ($\mu_1$and μ_2), and more than two populations means (μ_1, μ_2,... μ_k) which need tests of significance based on one sample, two samples and more than two samples respectively.

(i) One-sample Tests

The mental health care researcher may make hypothesis such as the mean age of onset of schizophrenic disorders is not equal to 30 years. The checking the plausibility of such a research hypothesis (H_1) may be carried out by comparing with the sample mean. The null and the alternate hypotheses are specified as,

$$H_0: \mu = \mu_0 \text{ against } H_1 : \mu \neq \mu_0$$

When the *sample size is large* (more than 30), the critical ratio and its distribution is given by,

$$Z = \frac{|\bar{x} - \mu_0|}{S/\sqrt{n}}$$

If the observed critical ratio is less than 1.96, then the null hypothesis is accepted. If the observed critical ratio is more then 1.96, then the null hypotheses is rejected at 5% level of significance and the alternate hypothesis is accepted with more than 95% confidence ($P<0.05$). If the observed critical ratio is more then 2.58, then the null hypothesis is rejected at 1% level of significance and the alternate hypothesis is accepted with more than 99% confidence ($P<0.01$). For example, let us make the hypothesis that the mean age of onset of schizophrenic disorders is not equal to 30 years. The null and the alternate hypotheses are specified as,

H_0: $\mu = 30$ years against H_1: $\mu \neq 30$ years.

The case history records of 124 schizophrenics registered at NIMHANS hospital revealed that the mean age of onset of the disorders is 31 years with a standard deviation of 11.78 years. The critical ratio is computed as,

$$Z = \frac{|31-30|}{11.78/\sqrt{124}} = \frac{1}{1.058} = 0.945$$

Since the calculated value of z is less then 1.96, the null hypothesis 'the mean age of onset of schizophrenic disorders is 30 years' is accepted with more than 95% confidence.

In case the *sample size is small*, the critical ratio cited above do not follow Z-distribution. The suitable critical ratio and its distribution is given by,

$$t = \frac{|\bar{x} - \mu_0|}{s/\sqrt{n}} \text{ with } df = (n-1)$$

Let us make an hypothesis that the mean age of onset of organic psychoses is not equal to 30 years. That is,

$$H_0: \mu = 30 \text{ against } H_1: \mu \neq 30 \text{ years}$$

The case history records of 14 organic psychotics registered at NIMHANS hospital revealed that the mean age of onset of this disorders is 39.1 years with an unbiased standard deviation of 14.202 years. Then,

$$t = \frac{|39.1-30|}{14.202/\sqrt{14}} = \frac{9.1}{3.796} = 2{,}397, \text{ with } df = (14-1) = 13$$

Since the calculated value is more than the table value of t with 13 degrees of freedom at 5% level of significance (2.16), the null hypothesis is rejected. Hence, the alternate hypothesis 'the mean age of onset of organic psychoses is not equal to 30 years, it is more than 30 years' is accepted with more than 95% confidence ($P<0.05$).

(ii) Two-sample Tests

The mental health researcher may make hypothesis such as the mean age of onset of schizophrenic disorders (μ_1) differs from the mean age of onset of neurotic disorders (μ_2). The null and alternate hypotheses are specified as,

$$H_0: \mu_1 = \mu_2 \text{ against } H_1: \mu_1 \neq \mu_2.$$

When the *sample size is large*, the critical ratio and its distribution is given by (with usual notations),

$$Z = \frac{|\bar{x} - \bar{y}|}{\sqrt{S_1^2/n_1 + S_2^2/n_2}}$$

The case history records of 124 schizophrenics revealed that the mean age of onset of these disorders is 31 years with a standard deviation of 11.78 years, and the case history records of 74 neurotic disorders patients revealed that the mean age of onset of these disorders is 27.2 years with a standard deviation of 11.56 years. Then,

$$Z = \frac{|31.0 - 27.21|}{\sqrt{(11.78^2 / 124) + (11.56^2 / 74)}} = \frac{3.79}{1.71} = 2.216$$

Since the calculated value of z is more than 1.96, the null hypothesis is rejected at 5% level of significance. Hence, the alternate hypothesis 'on the average, the onset of neurotic disorders is earlier than those of the schizophrenic disorders' is accepted with more than 95% confidence.

In case the *sample size is small*, the critical ratio cited above does not follow z-distribution. The suitable critical ratio and its distribution is given by,

$$t = \frac{|\bar{x} - \bar{y}|}{\sqrt{\frac{(\Sigma x^2 - n_1 \bar{x}^2) + (\Sigma y^2 - n_2 \bar{y}^2)}{(n_1 + n_2 - 2)} \times \left[\frac{1}{n_1} + \frac{1}{n_2}\right]}}$$

$$\text{with } df = (n_1 + n_2 - 2)$$

Let us suppose that the length of stay of five patients discharged on a particular day in an emergency ward are given by 4,2,3,0 and 6 days, and the length of stay of five patients discharged in another emergency ward are given by 7,4,1,3 and 5 days. The null hypothesis states that there is no significant difference between the mean length of stay of patients in these two wards. Then,

$$t = \frac{|3 - 4|}{\sqrt{\frac{(65 - 5 \times 3^2) + (100 - 5 \times 4^2)|}{(5+5-2)} \times (1/5 + 1/5)}} \quad \text{with } df = (5 + 5 - 2) = 8$$

$$= \frac{1}{\sqrt{2}} = 0.707$$

Since the calculated value is less than the table value of t with 8 degrees of freedom at 5% level of significance (2.31), the null hypothesis is accepted.

(iii) Paired-sample Test

Some times, the observed values are *related* such as the marks scores by a group of students in a test before conducting a training program and the marks scored by the same group of students after the training program. Generally, we make the hypotheses that the mean score is significantly increase after the training program. In this case, the critical ratio and its distribution is given by,

$$t = \frac{|\bar{d}|}{s_d / \sqrt{n}}$$

where d is the mean difference between the scores. The marks scored by 9 students in bio-statistics test before conducting a training program and after the training program are given below.

	Students								
Test	1	2	3	4	5	6	7	8	9
Before	5	8	10	14	3	9	15	11	7
After	11	13	12	14	4	10	16	9	11
Difference	6	5	2	0	1	1	1	-2	4

For this data, n = 9, $\bar{d} = 2.0$, $s_d = 2.55$. Then,

$$t = \frac{2.0}{2.55/\sqrt{9}} = \frac{2.0}{0.85} = 2.353$$

Since the observed value is more than the table value of t at 8 degrees of freedom (2.31), the null hypothesis is rejected at 5% level of significance. Hence, the alternate hypothesis 'the mean marks scored has significantly improved after conducting the test' is accepted with more than 95% confidence.

(b) Tests of Significance on Proportions

The researcher often involves in making certain hypotheses concerned with population proportions.

(i) One Sample Test

The mental health researcher may make an hypothesis such as the proportion of female registrations at NIMHANS hospital is not the same as that of the proportion of female patients of the psychiatric population in the Indian community (53.8%). The null and the alternate hypotheses are specified as,

$$H_0: P = P_0 \text{ against } H_1: P \neq P_0$$

When the sample size is large, the critical ratio and its distribution is specified as,

$$Z = \frac{|x - nP_0|}{\sqrt{nP_0Q_0}}$$

Where n is the sample size and x is the number of individuals in the samples posses the character. There are 197 (39.4%) females in a sample of 500 psychiatric registrations at NIMHANS hospital. The null hypothesis states that the proportion of female psychiatric registrations is the same as that of the psychiatric patients in the general population. Then,

$$Z = \frac{|197 - 500 \times 0.538|}{\sqrt{500 \times 0.538 \times 0.462}} = \frac{72}{11.148} = 6.459$$

Since the calculated value of z is more than 2.58, the null hypothesis is rejected at 1% level of significance. Hence, the alternate hypotheses 'the proportion of female registrations at NIMHANS hospital is less than that of the proportion of female patients among the psychiatric patients in the general population' is accepted with more than 99% confidence.

(ii) Two–sample Test

The mental health care researcher may involve in making hypotheses concerned with *two population proportions* (P_1 and P_2), such as the proportion of female registrations at Karnataka Institute of Mental Health (KIMH) Dharwad is different from the proportion of females in NIMHANS hospital Bangalore. The null and alternate hypotheses are specified as,

$$H_0: P_1 = P_2 \text{ against } H_1: P_1 \neq P_2$$

The samples are drawn independently from their respective populations and their sample proportions (p_1 and p_2) are to be compared. The critical ratio and its distribution is given by,

$$z = \frac{|p_1 - p_2|}{\sqrt{pq\,(1/n_1 + 1/n_2)}}$$

Where p is the pooled sample proportion obtained by combining both the samples. There are 30 females in a sample of 100 psychiatric registrations at KIMH hospital. The null hypothesis states that the proportion of female registrations at KIMH hospital is the same as that of NIMHANS hospital. Then,

$$Z = \frac{|0.300 - 0.394|}{\sqrt{0.378 \times 0.622 \times (1/100 + 1/500)}} = \frac{0.094}{0.053} = 1.770$$

Since the calculated value of z is less than 1.96, the null hypothesis is accepted at 5% level of significance.

(c) Test of Significance on Correlation Coefficient

The test of significance of *sample* correlation coefficient is of interest of mental health researcher whenever he computes and use the correlation coefficient based on the sample data. The null and alternate hypotheses are specified as,

$$H_0: \rho = 0 \text{ against } H_1: \rho \neq 0$$

Where ρ is the population correlation coefficient. The critical ratio and its distribution is given by,

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \quad \text{with } df = (n-2)$$

The correlation coefficient between the number of years of education and income per month of 35 psychiatric patients registered at NIMHANS hospital is calculated to be 0.658. The null hypothesis states that there is no significant correlation between the years of education and the income of psychiatric patients registered at NIMHANS hospital. Then,

$$t = \frac{0.658\ \sqrt{35-2}}{\sqrt{1-0.658^2}} = \frac{3.780}{0.753} = 5.04$$

Since the calculated value is more than the table value of t (3.646) with 33 degrees of freedom with 0.1% level of significance, the correlation coefficient is highly significant ($P<0.001$).

CHAPTER 12

ANALYSIS OF VARIANCE

The mental health care researcher often involves in making certain hypotheses concerned with *mean scores of more than two populations* ($\mu_1, \mu_2, \ldots \mu_k$), such as the mean ages of onset of organic psychoses, schizophrenic disorders, affective disorders and neurotic disorders are different. The null and the alternate hypotheses are specified as,

$$H_0: \mu_1 = \mu_2 = - - - = \mu_k \quad \text{against} \quad H_1: \mu_i \text{ are not equal.}$$

The test of significance of these hypotheses may be carried out by comparing the sample means of the characteristic from different populations. The statistical technique used here is known as *analysis of variance* (ANOVA). This technique consists in *breaking down* the total amount of variation in a set of data into two components, viz: the amount which can be attributed to specified causes and that amount which can be attributed to chance. It is possible to investigate any number of factors said to influence the dependent variable. The *principle* consists in estimating two population variances, one is based on between samples variance called the mean sum of squares between samples (MSS_B) and the other is based on within samples variance called mean sum of squares

within samples (MSS_w). The MSS_w is also known as mean sum of squares of errors. Then the two estimates are compared with the F-ratio. When the samples come from identical population, these two estimates of the population variance are comparable, any difference observed between them is expected to be within the range of their sampling errors. When the populations from which the samples are drawn have different means, the MSS_B is expected to provide a higher value.

Let us deal with *one factor* and investigate the differences among its various categories having numerous possible values. Let x_{ij} represents the j^{th} observation in the i^{th} group of the sample. The total sum of squares in a set of data is given by,

$$SS_T = \Sigma\Sigma (x_{ij} - \bar{x})^2 = \Sigma\Sigma x^2_{ij} - \frac{T^2}{N}$$

Where T is the total of the observations and N is the total number of observations. The sum of squares between groups is given by,

$$SS_B = \Sigma n_i (\bar{x}_i - \bar{x})^2 = \Sigma \frac{T_i^2}{n_i} - \frac{T^2}{N}$$

where T_i is the total of the observations, n_i is the number of observations and $\bar{x}_i$ is the mean of the observations in the i^{th} group. The sum of squares within groups is given by,

$$SS_w = \Sigma\Sigma (x_{ij} - \bar{x}_i)^2 = \Sigma\Sigma x_{ij}^2 - \Sigma \frac{T_i^2}{n_i}$$

Thus, $$SS_w = SS_T - SS_B$$

The mean sum of squares between groups is given by,

$$MSS_B = \frac{SS_B}{(k-1)}$$

The mean sum of squares within groups is given by,

$$MSS_W = \frac{SS_W}{(N-k)}$$

The critical ratio and its distribution is given by,

$$F = \frac{MSS_B}{MSS_w} \quad \text{with } df = k-1, N-k$$

It is conveniently present the computed values in a form of table known as ANOVA table as shown below.

Source of variation	df	SS	MSS	F-ratio
Between groups	k-1	SS_B	MSS_B	F
Within groups	N-k	SS_w	MSS_w	
Total	N-1	SS_T	—	—

Let us *suppose* that the scores obtained on an aptitude test of three groups of students of science, arts and commerce are as shown in the following table.

	Science students	Arts students	Commerce students	All students
	4	7	10	
	2	4	7	
	3	1	8	
	0	3	6	
	6	5	4	
n_i	5	5	5	15
T_i	15	20	35	70
$\bar{x}_i$	3	4	7	4.67
Σx^2	65	100	265	430

For this data, $SS_T = 430 - \frac{70^2}{15} = 430 - 326.67 = 103.33$

$$SS_B = \left(\frac{15^2}{5} + \frac{20^2}{5} + \frac{35^2}{5}\right) - 326.67 = 370.00 - 326.67 = 43.33$$

$$SS_E = 103.33 - 43.33 = 60.0$$

The ANOVA table is prepared as shown below.

Source of variation	df	SS	MSS	F-ratio
Between groups	2	43.33	21.67	4.33
Errors	12	60.00	5.00	–
Total	14	103.33	–	–

Since the calculated value of F is greater than the table value (3.89) with degrees of freedom of 2 and 12 at 5% level of significance, the null hypothesis is rejected. Hence, the alternate hypothesis 'the mean scores of the three groups of students are different' is accepted with more than 95% confidence ($P<0.05$). The results of the study may be presented in the form of table as shown below.

Statistics	Science group (5)	Arts group (5)	Commerce group (5)	Inference
Mean	3.0	4.0	7.0	$F = 4.33$; $df = 2,12$; $P<0.05$
SD	2.0	2.0	2.0	

(a) Post hoc Tests

A significant F-test in ANOVA does not mean that every group in the analysis is significantly different from every other group. Some times, the mental health care researchers are interested in carrying out the test

of significance of the difference in *pairs of means*. The tests applied after finding that the overall F is significant are called post hoc tests. Here the usual z-tests and the t-tests are cumbersome. The Scheffe method is the commonly used post hoc test. This method consists in calculating critical value of F to use as a standard against which to compare the differences in pairs of means. The critical value of F is given by,

$$F_{cv} = (k-1)\, F$$

Where k is the number of groups and F is that value needed to gain significance in the ANOVA. After calculating this, we find the Scheffe values for the differences in the pairs of means. The Scheffe value between the i^{th} group and the j^{th} group is given by (with usual notations),

$$\text{Scheffe Value} = \frac{(\bar{x}_i - \bar{x}_j)^2}{MSS_W (1/n_i + 1/n_j)}$$

The Scheffe value is compared with the F_{cv}. If the Scheffe value is greater than the F_{cv}, then we conclude that the two groups in the comparisons have significantly different means. In our example used to demonstrate ANOVA, the $F_{cv} = (3–1)\ (3.89) = 7.78$. To compare science students and arts students

$$\text{Scheffe Value} = \frac{(3-4)^2}{5(1/5+1/5)} = \frac{1}{2} = 0.50$$

To compare science students and commerce students,

$$\text{Scheffe Value} = \frac{(3-7)^2}{5(1/5+1/5)} = \frac{16}{2} = 8.00$$

To compare arts students and commerce students,

$$\text{Scheffe Value} = \frac{(4-7)^2}{5(1/5+1/5)} = \frac{9}{2} = 4.50$$

The Scheffe value for the comparison of science students and commerce students is larger than F_{cv}. Hence, the null hypothesis with respect to these two groups is rejected and the alternate hypothesis 'the commerce students are having significantly better aptitude than the science students' is accepted with more than 95% confidence ($P < 0.05$).

(b) Analysis of Covariance

In ANOVA, it is assumed that the observed values (y) are attributed to the treatments applied and not to any other causal circumstances. Some times, some concomitant variable (x) is correlated with the dependent variable (y). The marks scored in psychiatry by a group of multipurpose health workers after conducting a training program may be related to the marks scored by the group before conducting the training program. In such situations, the researcher should use the statistical technique of *analysis of covariance* (ANCOVA) for valid comparison of treatment effects. This technique consists in subtracting from each individual score (y_i) that portion which is predictable from the concomitant variable, obtaining adjusted scores and carrying out the usual ANOVA. The use of *regression analysis* required the loss of one degree of freedom. In ANCOVA, the adjusted sum of squares are obtained by calculating the sum of products. The total sum of products (SP_T), the between groups sum of products (SP_B), and the within groups sum of products (SP_W) are given by,

$$SP_T = \sum\sum x_{ij}\, y_{ij} - \frac{T_x T_y}{N}$$

$$SP_B = \sum \frac{T_{xi}\, T_{yi}}{n_i} - \frac{T_x T_y}{N}$$

$$SP_W = SP_T - SP_B$$

The adjusted total sum of squares ($SS`_{YT}$), the adjusted within groups sum of squares ($SS`_{yw}$) and the adjusted between groups sum of squares ($SS`_{yB}$), are given by,

$$SS`_{YT} = SS_{YT} - \frac{SP^2_T}{SS_{XT}}$$

$$SS`_{YW} = SS_{YW} - \frac{SP^2_W}{SS_{XW}}$$

$$SS`_{YB} = SS`_{YT} - SS`_{YW}$$

It is convenient to prepare an ANCOVA table as shown below.

Source of variation	df	SS_X	SP	SS_Y	SS'_Y	MSS'_Y	F-ratio
Between groups	(k–1)	SS_{XB}	SP_B	SS_{YB}	SS'_{YB}	MSS'_{YB}	F
Within groups	(N–k–1)	SS_{XW}	SP_W	SS_{YW}	SS'_{YW}	MSS'_{YW}	–
Total	(N–2)	SS_{XT}	SP_T	SS_{YT}	SS'_{YT}	–	–

Let us *suppose* that the marks scored in psychiatry (x) by three groups of multipurpose workers before conducting a training program and the marks scored by the groups (y) after conducting the training program are as shown in the following table.

	Group I		Group II		Group III		Total	
	x	y	x	y	x	y	x	y
	4	4	6	7	8	10		
	2	2	5	4	6	7		
	1	3	3	1	9	8		
	0	0	3	3	8	6		
	3	6	8	5	4	4		
Sum	10	15	25	20	35	35	70	70
Mean	2	3	5	4	7	7	4.67	4.67
SS	30	65	143	100	261	265	434	430

For this data,

$$SS_{XT} = 434 - \frac{70^2}{15} = 434 - 326.67 = 107.33$$

$$SS_{XB} = \left[\frac{10^2}{5} + \frac{25^2}{5} + \frac{35^2}{5}\right] = 390 - 326.67 = 63.33$$

$$SS_{XE} = 107.33 - 63.33 = 44.00$$

and $$SP_T = 413 - \frac{70\times70}{15} = 86.33$$

$$SP_B = \left[\frac{10\times15}{5} + \frac{25\times20}{5} + \frac{35\times35}{5}\right] - \frac{70\times70}{15} = 48.33$$

$$SP_E = 86.33 - 48.33 = 38.0$$

and $$SS_{YT} = 430 - \frac{70^2}{15} = 430\text{-}326.67 = 103.33$$

$$SS_{YB} = \left[\frac{15^2}{5} + \frac{20^2}{5} + \frac{35^2}{5}\right] - 326.67 = 370 - 326.67 = 43.33$$

$$SS_{YE} = 103.33 - 43.33 = 60.$$

and $$SS`_{YT} = 103.33 - \frac{86.33^2}{107.33} = 33.89$$

$$SS`_{YE} = 60 - \frac{38^2}{44} = 27.18$$

Hence $$SS`_{YB} = 33.89 - 27.18 = 6.71$$

The MSS_Y' are obtained by dividing the SS'_y by their respective degrees of freedom. The computed values are presented in the following ANCOVA table.

Source	df	SS_X	SP	SS_Y	SS'_Y	MSS'_Y	F-ratio
Between groups	2	63.33	48.33	43.33	6.71	3.355	1.358
Errors	11	44.00	38.00	60.00	27.18	2.471	–
Total	13	107.33	86.33	103.33	33.89	–	–

Since the calculated value is less than the table value of F (4.00) with degrees of freedom of 2 and 11 at 5% level of significance, the null hypothesis 'the mean scores of the three groups after the treatment are the same' is accepted.

CHAPTER 13

NON-PARAMETRIC TESTS OF SIGNIFICANCE

The non-parametric tests of significance are the methods of inference in which there are *no assumptions* whatsoever have been made about the nature, shape or form of the populations from which the data are obtained. Hence, these tests are often called as *distribution free methods, categorical data analysis, order statistics and ranking tests.* These tests are typically much easier to learn and to apply. Some of these tests are often used as computational short-cuts. They give more preferable results when the sample size is small. If all the assumptions are met in the data, then these tests are wasteful of data. The degree of wastefulness is measured in terms of *power efficiency* of the test. Most of these tests do not make use of the absolute values of measurement. There are several non-parametric tests which may be broadly classified according to the levels of measurement of variables, number and nature of samples, and the purpose of the test of significance.

(a) Chi-square Test of Significance

The chi-square tests are *commonly* used non-parametric tests of significance in mental health care research field. These tests are applicable when the variables are measured in or reduced to nominal level and the samples are drawn independently. They may be used as tests of goodness

of fit of distributions, as test of significance of association of attributes and as tests of significance of differences / trends of proportions. The critical ratio and its distribution is given by,

$$\chi^2 = \Sigma\Sigma \frac{(o_{ij} - e_{ij})^2}{e_{ij}} = \Sigma\Sigma \frac{o_{ij}^2}{e_{ij}} - n \qquad \text{with df} = (r-1)(c-1)$$

Where o_{ij} and e_{ij} are the observed and expected frequencies respectively of the cell in the i^{th} column and j^{th} row of the contingency table. Under null hypothesis, the expected frequency e_{ij} is given by,

$$e_{ij} = \frac{n_{i.}\, n_{.j}}{n}$$

Where $n_{i.}$ is the total of the frequencies in the i^{th} column (marginal total), $n_{.j}$ is the total of the frequencies in the j^{th} row, and n is the total of the frequencies in the contingency table. The expected frequencies should be sufficiently large. In case of 2×2 contingency table, each expected frequency should be 5 or large. In case of more classifications, no more than 20% of expected frequencies should be smaller than 5. The degrees of freedom (df) for chi-square dealing with contingency table is given by,

$$\text{df} = (\text{number of rows} - 1) \times (\text{number of columns} - 1)$$

In contingency tables concerned with classification of data according to *single variable,* the null hypothesis states that the observed frequencies are according to certain assumptions. The critical ratio and its distribution is given by,

$$\chi^2 = \Sigma \frac{(o_i - e_i)^2}{e_i} = \Sigma \frac{o_i^2}{e_i} - n \qquad \text{with df} = (k-1)$$

Where o_i and e_i are the observed and expected frequencies respectively of the i^{th} category of the contingency table. Let us make an hypothesis that the number of psychiatric registrations at NIMHANS hospital on each working day are significantly different . The number of registrations on each of the six working days during the first week of April 1999 at NIMHANS hospital, are shown in the following table . The expected number of registrations per working day under the null hypothesis is obtained by dividing the total of the frequencies by the number of working days.

Week day	Mon	Tue	Wed	Thu	Fri	Sat	Total
o_i	24	26	23	45	25	13	156
e_i	26	26	26	26	26	26	156
o_i^2/e_i	22.15	26.00	20.35	77.88	24.04	6.50	176.92

Here, $\chi^2 = 176.92 - 156 = 20.92$ with df = (6-1) = 5

Since the calculated value is greater than the table value of chi-square with 5 degrees of freedom at 1% level of significance (15.09), the null hypothesis is rejected. Hence, the alternate hypothesis 'the number of psychiatric registrations on each working day at NIMHANS hospital are different' is accepted with 99% confidence ($P< 0.01$).

A formula for 2×2 *contingency table* is derived as given below.

$$\chi^2 = \frac{n(|ad - bc| - \frac{n}{2})^2}{(a+b)(c+d)(a+c)(b+d)} \quad \text{with df} = 1$$

Where the format of the contingency table is shown below:

Variable I	Variable II		Total
	Category 1	Category 2	
Category 1	a	b	(a+b)
Category 2	c	d	(c+d)
Total	(a+c)	(b+d)	n

Let us make an hypothesis that gender and the type of service (out-patients service and in-patient service) of psychiatric registrations at NIMHANS hospital are associated. The classification according to gender and type of service of the 500 consecutively registered psychiatric patients of NIMHANS hospital is as given below.

Gender	In-patients	Out-patients	Total registrations
Male	111(70.2)	192(56.1)	303(60.6)
Female	47(29.8)	150(43.9)	197(39.4)
Total	158(100.0)	342(100.0)	500(100.0)

For this data,

$$\chi^2 = \frac{500(|111\times150 - 192\times47| - 250)^2}{303\times197\times158\times342} = 8.434 \quad \text{with df} = 1$$

Since the calculated value is greater than the table value of χ^2 with 1 degree of freedom at 1% level of significance (6.63%), the null hypothesis that the type of service is not associated with gender is rejected. Hence the alternate hypothesis 'the gender and type of service are associated' is accepted with 99% confidence interval.

A formula for 2 × k *contingency table* is derived as given below.

$$\chi^2 = \frac{z^2}{xy}\left[\frac{x_1^2}{z_1} + \frac{x_2^2}{z_2} + \ldots\ldots + \frac{x_k^2}{z_k}\right] - \frac{zx}{y} \quad \text{with df} = (k\text{-}1)$$

Where the format for the contingency table is given by,

Variable I	Variable II Categories						Total
	1	2	–	–	–	k	
Category 1	x_1	x_2	–	–	–	x_k	x
Category 2	y_1	y_2	–	–	–	y_k	y
Total	z_1	z_2	–	–	–	z_k	z

A classification according to religion and the type of service of the 500 psychiatric patients is as given in the following table.

Type of service	Hindu	Muslim	Christian	Total
In-patients	133(32.8)	14(21.2)	11(39.3)	158(31.6)
Out-patients	273(67.2)	52(78.8)	17(60.7)	342(68.4)
Total registrations	406(100.0)	66(100.0)	28(100.0)	500(100.0)

For this data,

$$\chi^2 = \frac{500^2}{158\times342}\left[\frac{133^2}{406}+\frac{14^2}{66}+\frac{11^2}{28}\right]-\frac{500\times158}{342} = 4.39 \quad \text{with df} = (3-1) = 2$$

Since the calculated value is less than the table value of χ^2 with 2 degrees of freedom at 5% level of significance (5.99), the null hypothesis that the type of service is not associated with religion is accepted.

In case of *r* ×*k contingency table*, the critical ratio and its distribution is given by,

$$\chi^2 = \Sigma\Sigma \frac{o_{ij}^2}{e_{ij}} - n \qquad \text{with df} = (r-1)\,(k-1)$$

The classification according to religion and diagnostic groups of schizophrenic disorders (Sch), affective disorders (AD) and neurotic disorders (ND) of 337 psychiatric patients is given in the following table. The expected frequencies are calculated as shown in brackets.

Religion	Sch	AD	ND	Total
Hindu	109 (102.66)	114 (115.08)	56 (61.26)	279
Muslim	8 (16.19)	22 (18.15)	14 (9.66)	44
Christian	7 (5.15)	3 (5.77)	4 (3.08)	14
Total	124	139	74	337

For this data,

$$\chi^2 = \left[\frac{109^2}{102.66} + \frac{114^2}{115.08} + - - - + \frac{4^2}{3.08} \right] - 337$$

$= 347.034 - 337 = 10.034$ with df = (3-1)(3-1)=4

Since the calculated value is greater than the table value of chi-square with 4 degrees of freedom at 5% level of significance (9.49), the null hypothesis is rejected. The alternate hypothesis 'the religion and the diagnostic groups are associated' is accepted with more than 95% confidence ($P < 0.05$).

(b) Other Non-parametric Tests of Significance

The chi-square tests are used as one-sample tests, as two-independent sample tests and as k-independent sample tests as illustrated above. Besides the chi-square tests, there are several non-parametric tests of significance which may be classified according to the levels of measurement of variables and types of samples. Some of the important non-parametric tests are described below.

(i) One-sample Tests

The *binomial test* is applied when there are two categories in the classification. The test is uniquely used when the sample size is so small that the chi-square test is inapplicable. The binomial test may be used

as test of goodness of fit when the null hypothesis states that the frequencies are according to binomial distribution. The probability of obtaining x objects in one category and (n-x) objects in other category is given by,

$$P(x) = {}^{n}c_{x} p^{x} q^{n-x}$$

The probability of obtaining the observed value or values even more extreme is given by,

$$P(x \leq k) = \Sigma\, {}^{n}c_{x} p^{x} q^{n-x}$$

The prevalence of mental and behavioural disorders in India is estimated to be 58.2 per one thousand population. We are interested in knowing the probability of obtaining two or more psychiatric patients in a family of five members in India. We know that,

$$P = 1 - P(0) - P(1)$$

To determine these probabilities, we have

$$P(0) = \frac{5!}{0!5!} (0.0582)^{0} (0.9418)^{5} = 0.741$$

$$P(1) = \frac{5!}{1!4!} (0.0582)^{1} (0.9418)^{4} = 0.229$$

$$P = 1 - 0.741 - 0.229 = 1 - 0.97 = 0.03$$

Thus the probability of obtaining two or more psychiatric patients in a family of five members in India is 0.03 or 3%.

The ***run test*** is applied when the null hypothesis says that the sequence of events in a sample is at random. It is based on the number of runs (r) which a sample exhibits, the number of objects in one category (n_1)

and the number of objects in another category (n_2). A run is defined as a success of identical symbols which are followed and proceeded by different symbols or by no symbols at all. Too many or too small number of runs do not support randomness . When the sample size is large, the critical ratio and its distribution is given by,

$$z = \frac{|r - \mu_r|}{\sigma_r}$$

where $\mu_r = \frac{2n_1n_2}{(n_1+n_2)} + 1$ and $\sigma_r = \sqrt{\frac{2n_1n_2(2n_1n_2-n_1-n_2)}{(n_1+n_2)^2\,(n_1+n_2-1)}}$

The psychiatric registrations of NIMHANS hospital on a particular day has occurred in the following order of males and females as shown below.

FFF MM FF MMM F M F MM F M F M FF M F MMMM F M F MMMM F M F MMMM F MM

This sample of scores begin with a run of 3 females. A run of 2 males follows and so on. There are 26 runs (r) for 27 males (n_1) and 17 females (n_2). Hence

$$\mu_r = \frac{2 \times 27 \times 17}{(27 + 17)} + 1 = 21.864$$

$$\sigma_r = \sqrt{\frac{(2 \times 27 \times 17)\,(2 \times 27 \times 17 - 27 - 17)}{(27 + 17)^2\,(27 + 17 - 1)}} = \sqrt{\frac{918\,(918-44)}{44^2\,(44 - 1)}} = 3.104$$

$$Z = \frac{|26 - 21.864|}{3.104} = \frac{4.136}{3.104} = 1.332$$

Since the calculated value is less than 1.96, we accept the null hypothesis that the sequence of males and females is at random.

(b) Two-sample Tests

The ***Fisher exact probability test*** is used when the data is in 2×2 contingency table and the expected frequency in a cell is small. It is used to test whether two samples represent population which differ in location. The exact probability of observing a particular set of frequencies in a 2×2 contingency table when the marginal totals are regarded as fixed is given by,

$$P(x) = \frac{(a+b)!(c+d)\,!(a+c)!(b+d)!}{n\,!\,a\,!\,b\,!c\,!\,d\,!}$$

Where a, b, c and d are the frequencies of the contingency table. Let us suppose that x is the smallest frequency. Then the probability of occurrence of or even more extreme occurrence is given by,

$$P = P(x) + P\,(x\text{-}1) + \ldots + P(0)$$

A census of long-stay patients (LSP) in all the 36 government mental hospitals in India as on the first july 1999, revealed that 53% of the patients are males. There were 13 hospitals which had high percentage of males when compared with the national indicator. Further, six of the hospitals had high percentage of criminal record patients when compared with the national indicator of 4% as shown in the following table.

Proportion of males	Proportion of criminal record patients		Total hospitals
	High	Low	
High	6	7	13
Low	0	23	23
Total	6	30	36

For this data, P = P (0) and

$$P(0) = \frac{13!\ 23!\ 6!\ 30!}{36!\ 6!\ 7!\ 0!\ 23!} = 0.00088$$

Since the calculated exact P value is less than 0.1%, the association between male gender and criminal records among long stay patients is highly significant.

The ***median test*** may be used when the null hypothesis states that the two groups are from populations with the same median This method is applicable whenever the scores for the two groups are in at least an ordinal scale. To perform the median test, we first determine the median score for the combined group. Then we dichotomize both sets of scores at that combined median and cast these data as shown in the following 2 × 2 table.

Category	Group I	Group II	Total
Above median	a	b	(a+b)
Below median	c	d	(c+d)
Total	(a+c)	(b+d)	n

Then the procedure is to apply the chi-square test by considering it as 2×2 contingency table. The median duration of illness of 500 psychiatric patients registered at NIMHANS hospital is found to be one year and two months. The classification of in-patients and out-patients according to the combined median is given below.

Duration of Illness	In-patients	Out-patients	Total registrations
Above median	91(57.6)	159(46.5)	250 (50.0)
Below median	67(42.4)	183(53.5)	250 (50.0)
Total	158(100.0)	342(100.0)	500 (100.0)

For this data

$$\chi^2 = \frac{500(|91 \times 183 - 159 \times 67| - 250)^2}{250 \times 250 \times 158 \times 342} = 4.895 \qquad \text{with df} = 1$$

Since the calculated value is greater than the table value of χ^2 with 1 degree of freedom at 5% level of significance (3.84), the null hypothesis is rejected. Hence the alternate hypothesis 'the median duration of illness of in-patients is greater than that of the out-patients at NIMHANS Hospital' is accepted with more than 95% confidence ($P < 0.05$).

(iii) Two-related Sample Tests

The *McNemar test* is used when the researcher involves in hypothesis that there is change after the treatment. This test of significance is used when both the variables under study are measured in nominal scale. The test is equivalent to a binomial test with $P = Q = 0.5$ and n is the number of changes. Hence the researcher sets up a four-fold table of frequencies to represent first and second sets of responses from the same individuals. A typical table is given below in which fail and pass are used to signify different responses. The critical ratio and its distribution is given by,

$$\chi^2 = \frac{(|A - D| - 1)^2}{(A + D)} \qquad \text{with df} = 1$$

Where the data format is illustrated below.

		After	
		Pass	Fail
Before	Fail	A	B
	Pass	C	D

The trends of average bed occupancy from 1977 to 1983 and the trends from 1984 to 1999 in government mental hospitals in India are as shown below.

Trends during 1977-83	Trends during 1984-99		Total
	Increasing	Decreasing	
Decreasing	4	14	18
Increasing	2	16	18
Totals	6	30	36

For this data,

$$\chi^2 = \frac{(|4-16|-1)^2}{(4+16)} = \frac{11^2}{20} = 6.05 \qquad \text{with df} = 1$$

Since the calculated value is grater than the table value of chi-square with 1 degree of freedom at 5% level of significance (3.84), the null hypothesis is rejected. Hence the alternate hypothesis 'there is a significant change in the rate of reduction of average bed occupancy during the period from 1984 to 1999 at government mental hospitals in India' is accepted with more than 95% confidence ($P < 0.05$).

The ***sign test*** is applied when the data on a variable has underlying continuity but which can be measured in only a very gross way. It uses plus and minus signs rather than quantitative measures as its data The method consists in focusing on the direction of the differences between every pair of observation and record plus sign when the after value is less than the value before giving the treatment, and minus sign otherwise. For small samples, the investigator may consult the binomial probabilities. In case of large samples, the critical ratio and its distribution is given by,

$$Z = \frac{|x - \frac{1}{2} n|}{\frac{1}{2}\sqrt{n}}$$

Where x is the number of plus signs and n is the number of observations. The load of long stay patients at 31 (86%) of 36 government mental hospitals were in decreasing trend during the period from 1999 to 2001. Hence,

$$Z = \frac{|31 - 0.5 \times 36|}{0.5\sqrt{36}} = 13/3 = 4.33$$

Since the calculated value is greater than 2.58, the alternate hypothesis stating that there is a significant reduction in the load of chronic patients at government mental hospitals in India during the period of 3 years from 1999 to 2001 is accepted.

(iv) k-sample Tests

In k-independent sample case, the chi-square test is used when the variable is measured in nominal scale.

In case of ordinal scale, the extended median test and the Kruskal-Wallis H-test are commonly used. The *Kruskal-Wallis test* is found more efficient than that of the extended median test because it uses more of the information in the observations. It converts the scores as ranks whereas the median test converts them simply to either plus or minus. In this test, each of the n observations are replaced by ranks. That is, all of the scores from all of the k samples combined are ranked in a single series. The smallest score is replaced by rank 1, the next to smallest by rank 2, and the largest by rank n. Here n is the total number of independent observations in the k-samples. Then the sum of the ranks in each sample (column) is found. The test determines whether these

sum of ranks are so desperate that they are not likely to have come from samples which were all drawn from the same population. The statistics denoted by H is given by,

$$H = \frac{12}{n(n+1)} \sum_{j=1}^{k} \frac{R_j^2}{n_j} - 3(n+1) \qquad \text{with } df = (k-1)$$

where k is the number of samples, n_j is the number of cases in the j^{th} sample, n is the number of cases in all samples combined and R_j is the sum of ranks in the j^{th} sample (column). The statistic H is distributed approximately as chi-square with (k-1) degrees of freedom.

The number of long-stay patients in government mental hospitals in India according to four broad regions are as shown in the following table.

Southern region (8)	Northern & Central region (9)	Western region (9)	Eastern region (10)
71	90	80	191
20	314	247	99
57	56	131	175
84	44	8	60
305	38	7	2
61	107	835	10
381	142	1848	79
692	133	470	89
	192	78	103
			9

If we rank the hospitals according to the load of long-stay patients from lowest to highest, we obtain the ranks as shown in the following table.

Southern region (8)	Northern & Central region (9)	Western region (9)	Eastern region (10)
13	19	16	27
6	31	29	20
10	9	23	26
17	8	3	11
30	7	2	1
12	22	35	5
32	25	36	15
34	24	33	18
	28	14	21
			4
Total: 154	173	191	148

For this data,

$$H = \frac{12}{36(36+1)}\left[\frac{154^2}{8} + \frac{173^2}{9} + \frac{191^2}{9} + \frac{148^2}{10}\right] - 3(36+1)$$

$$= 112.917 - 111 = 1.917 \qquad \text{with df} = (4-1) = 3$$

Since the calculated value is less than the table value of chi-square with 3 degrees of freedom at 5% level of significance (7.82), the null hypothesis 'there is no significant difference in the load of long-stay patients of government mental hospitals in various regions of India' is accepted.

(v) k-related Sample Tests

The *Friedman test* of significance is used when the researcher has the hypothesis that the k-related samples have come from different populations with respect to mean ranks. The test is applicable when the

measurement of the variable is at least in an ordinal scale. Since the k-samples are matched, the number of cases is the same in each of the samples. The data are cast in a two-way table having n rows and k columns. The rows represent the various subjects or matched sets of subjects, and the columns represent the various conditions. The data of the tests are ranks. The scores in each row are ranked separately. With k conditions being studied, the ranks in any row ranks from 1 to k. The critical ratio and its distribution is given by,

$$\chi_r^2 = \frac{12}{n\,k\,(k+1)} \sum_{j=1}^{k} R_j^2 - 3\,n\,(k+1) \qquad \text{with } df = (k-1)$$

Where n is the number of rows, k is the number of columns and R_j is the sum of ranks of the j^{th} column The average bed occupancy of 12 government mental hospitals during 1977, 1983 and 1999 are given below.

Hospital at	1977	1983	1999
1. Tezpur	938	952	353
2. Ranchi	1623	1475	543
3. Ranchi CIP	469	518	360
4. Delhi	363	412	140
5. Panaji	228	243	150
6. Bangalore	601	457	364
7. Trissur	345	518	382
8. Nagpur	1179	980	786
9. Ratnagiri	321	395	183
10.Amritsar	790	787	415
11.Chennai	1647	1358	1657
12.Varanasi	294	243	258

To perform the Friedman test on these data, we first rank the scores in each row we may give the lowest score in each row the ranks of 1, the next lowest score in each row the rank 2, etc. By doing this we obtain the data as shown in the following table.

Hospital at	1977	1983	1999
1. Tezpur	2	3	1
2. Ranchi	3	2	1
3. Ranchi CIP	2	3	1
4. Delhi	2	3	1
5. Panaji	2	3	1
6. Bangalore	3	2	1
7. Trissur	1	3	2
8. Nagpur	3	2	1
9. Ratnagiri	2	3	1
10. Amritsar	3	2	1
11. Chennai	2	1	3
12. Varanasi	3	1	2
R_j	28	28	16

For this data, n = 12 and k = 3, Hence

$$\chi_r^2 = \frac{12}{12 \times 3\,(3+1)}\,(28^2 + 28^2 + 16^2) - 3 \times 12\,(3+1) \qquad \text{with df} = (k-1)$$

$$= (1/12)\,(1824) - 144 = 152 - 144 = 8.00 \qquad \text{with df} = 2$$

Since the calculated value is greater than the table value of chi-square with 2 degrees of freedom at 5% level of significance (5.99), we reject the null hypothesis . Hence the alternate hypothesis ' the related samples have come from different populations with respect to mean ranks' is accepted with more than 95% confidence ($P < 0.05$).

CHAPTER 14

FURTHER ANALYSIS OF CONTINGENCY TABLES

The identification of frequencies responsible for a significant chi-square value, and fitting models and estimate parameters in the models are also *important* approaches in the analysis of contingency tables.

(a) Tests of Significance of Individual Frequencies

This approach consists in computation and examination of *adjusted standardized residual* for each cell in the contingency table. The adjusted standardized residual of the frequency in the i^{th} row and j^{th} column of the contingency table is given by,

$$Z_{ij} = \frac{\in_{ij}}{\sqrt{Var(\in_{ij})}}$$

Where $\in_{ij}$ is the *standardized residual* of the frequency in the ij^{th} cell. It is given by,

$$\in_{ij} = \frac{(o_{ij} - e_{ij})}{\sqrt{e_{ij}}}$$

Where o_{ij} and e_{ij} are the observed and the expected frequencies respectively of the ij^{th} cell. Under the null hypothesis, the e_{ij} is obtained by dividing the product of the marginal totals of the ij^{th} cell by the total of the frequencies. An estimate of the *variance* of $\in_{ij}$ is given by,

$$Var(\in_{ij}) = (1-p_{i.})\,(1-p_{.j})$$

where $p_{i.}$ and $p_{.j}$ are the proportions of the marginal totals of the i^{th} row and the j^{th} column respectively of the ij^{th} cell. When the variables forming the contingency table are independent, the z_{ij} are approximately normally distributed with mean 0 and standard deviation of 1.

The classification according to religion and type of service of 500 psychiatric patients of NIMHANS hospital is presented in the following table. The expected frequencies are computed as shown in the parentheses of the table. For example, the expected number of Hindu in-patients is obtained as (406×158)/500 = 128.30.

Religion	In-patients	Out-patients	Total
Hindu	133 (128.30)	273 (277.71)	406 (81.2)
Muslim	14 (20.85)	52 (45.14)	66 (13.2)
Christian	11 (8.85)	17 (19.15)	28 (5.6)
Total	158	342	500

$\chi^2 = 4.390$, df = 2, not significant (P>0.05)

The standardized residuals are computed as presented in the following table. For example, the standardized residual for Hindu in-patients is obtained as $(133-128.30)/\sqrt{128.30} = 0.415$

Religion	In-patients	Out-patients
Hindu	0.415	–0.282
Muslim	–1.501	1.020
Christian	0.724	–0.492

The variances of the standardized residuals are computed as presented in the following table. For example, the variance of the standardized

residual of Hindu in-patients is obtained as (1–406/500) (1–158/500) = (1–0.812) (1–0.316) = 0.188 × 0.684 = 0.129

Religion	Variances		$1-p_{i.}$
	In-patients	Out-patients	
Hindu	0.129	0.059	0.188
Muslim	0.594	0.274	0.868
Christian	0.646	0.298	0.944
$1-p_{.j}$	0.684	0.316	–

The adjusted standardized residuals are computed as shown in the following table. For example, the adjusted standardized residual for Hindu in-patients frequency is obtained as, $0.415/\sqrt{0.129} = 1.155$.

Religion	In-patients	Out-patients
Hindu	1.155	–1.155
Muslim	–1.948	1.948
Christian	0.901	–0.901

It can be noted that none of the z_{ij} of the frequencies is more than 1.96 and hence not significant. However the proportion of frequency of Muslim out-patients is nearly significantly high at 5% level of significance.

(b) Log-linear Models

The term model refers to some theory or conceptual framework about the observations. The parameters in the model represent the effects that particular variables or combinations of variables have in determining the observed values. Such an approach is common in regression analysis and analysis of variance. Most common are linear models which postulates that the expected values of the observations are given by a

linear combination of a number of parameters. Techniques such as maximum likelihood and least squares may be used in estimating the parameters. Then the estimated parameter values are used in identifying variables which have greatest importance in determining the observed values. The ANOVA term 'interaction' is used instead of the term 'association' for describing a relationship between the qualitative variables forming a contingency table. We shall speak of first-order interactions between pairs of variables, second-order interactions between triplets of variables, and so on. The major advantages of these techniques are that they provide a systematic approach to the analysis of complex multidimentional tables, provide estimates of the *magnitude of effects* of interest and consequently they allow the relative importance of different effects to be judged.

Let us consider how the type of model used in the analysis of variance of quantitative data can arise for contingency table data. Let us deal with the *two-dimensional table and the hypothesis of independence*. That is no first-order interaction between the two variables. It is specified by,

$$p_{ij} = p_{i.}\ p_{.j}$$

This relationship specifies a model for the data. It is that in the population the probability of an observation falling in the ij^{th} cell of the table is simply the product of the marginal probabilities. We wish to ask how this model could be rearranged so that p_{ij} can be expressed as the sum of the marginal probabilities or some function of them. By taking the natural logarithms of the above expression, we found,

$$\ln p_{ij} = \ln p_{i.} + \ln p_{.j}$$

$$e_{ij} = np_{ij} = n\ p_{i.}\ p_{.j}$$

$$= n \frac{e_{i.}\, e_{.j}}{n\ n} = \frac{e_{i.}\, e_{.j}}{n}$$

We have, $\ln e_{ij} = \ln e_{i.} + \ln e_{.j} - \ln n$

This equation may be rewritten in a form reminiscent of the models used in ANOVA, namely:

$$\ln e_{ij} = u + u_1(i) + u_2(j)$$

This is the linear model for the logarithms of the frequencies or what is generally known as a *log-linear model.* Here,

$$u = \frac{\Sigma\Sigma \ln e_{ij}}{rc}$$

$$u_1(i) = \frac{\sum_{j=1}^{c} \ln e_{ij}}{c} - u$$

$$u_2(j) = \frac{\sum_{i=i}^{r} \ln e_{ij}}{r} - u$$

Where u is the overall mean effect, $u_1(i)$ is the main effect of the i^{th} category of variable 1 and $u_2(j)$ is the main effect of the j^{th} category of variable 2. The numerical subscripts of the parameters denote the particular variables involved, and the alphabetic subscripts denote the categories of these variables in the same order.

Let us compute the main effects parameters of the model, $\ln e_{ij} = u + u_1(i) + u_2(j)$ for the religion and type of service data. The $\ln e_{ij}$ values are as shown in the following table

Religion	In-patients	Out-patients	Total
Hindu	4.854	5.627	10.481
Muslim	3.037	3.810	6.847
Christian	2.180	2.952	5.132
Total	10.071	12.389	22.460

The estimate of the main effect parameter $u_1(1)$ is obtained by,

$$u_1(1) = \frac{1}{2}(10.481) - \frac{1}{6}(4.854 + - - - + 2.952) \quad = 5.241 - 3.744 \quad = 1.497$$

The first set of estimate *main effects* are as shown in the following table

	Religion	Type of service
	$u_1(1) = 1.497$	$u_2(1) = -0.386$
Category	$u_1(2) = -0.320$	$u_2(2) = 0.386$
	$u_1(3) = -1.177$	

The sum of the estimates for each variable is zero. The size of the effects simply reflects the size of the marginal totals. That is of the parameters $u_1(i)$, $u_1(1)$ is the largest since the first category (Hindu) of variable 1 (Religion) has the largest marginal total among those of the variable. Similarly of the parameters $u_2(j)$, $u_2(2)$ is the largest.

When the two variables *are not independent*, the log-linear model is given by,

$$\ln e_{ij} = u + u_1(i) + u_2(j) + u_{12}(ij)$$

Where $u_{12}(ij)$ represents the interaction effect between levels i and j of variables 1 and 2 respectively. Interaction effects are measured as deviations and we have,

$$\sum_{j=1}^{c} u_{12}(ij) = 0 \quad \text{and} \quad \sum_{i=1}^{r} u_{12}(ij) = 0$$

Estimation of the interaction effects would be useful in identifying those categories responsible for any departure from independence.

CHAPTER 15

EXPERIMENTAL STUDIES

The mental health care researcher with some predefined objectives may adopt or construct suitable study design depending upon the need and availability. The studies based on interaction with the study subjects may be classified as experimental or observational. In experimental studies, the investigator manipulates the independent variable, assigning subjects to experimental and control groups randomly, and then observe the outcome. Hence the experimenter has *formed the population* purposefully in a specified way. There are experimental errors which includes all types of extraneous variation due to intrinsic variability of experimental material, lack of uniformity in the methodology of conducting the experiment and lack of representation of the sample to the population.

An experimenter resorts to the principle of *replication* in order to average out the influence of the chance factors on different experimental units. Thus the replication of the experiment results in more reliable estimates than is possible with a single observation. Replication reduces experimental error and thus enables us to obtain more precise estimates.

of the treatment effects. The precision of the experiment is proportional to the square root of the number of replications. The adequate number of replications for various treatments in an experiment depends upon the knowledge of the variability of the experimental material. Replication provides an estimate of the experimental error without which it is not possible to test the significance of the difference between treatment effects or determine the level of the confidence interval. The experimenter resorts to the principle of *randomization* in order to provide a logical basis for the validity of the statistical tests of significance. Randomization insures that different treatments are subjected to equal environmental effect on the average. Randomization eliminates bias in any form. It equalizes factors of variation over which we have no control. The experimenter may resort to the principle of *local control* in order to reduce the experimental error by dividing the heterogeneous experimental material into blocks consists of homogeneous units.

(a) Informal Designs

The experimental designs may be formal or informal. The informal experimental designs are those designs that normally use a less sophisticated form of analysis based on differences in magnitudes, such as before-and-after without control design, after-only with control design, and before-and-after with control designs. In *before-and-after without control design*, a single test group is selected and the dependent variable is measured before the introduction of the treatment. The treatment is then introduced and the dependent variable is measured again after the treatment has been introduced. The effect of the treatment would be equal to the level of the phenomenon after the treatment minus the level

of the phenomenon before the treatment. The main difficulty here is that with the passage of time considerable extraneous variations may be there in the treatment effect.

In *after-only with control design*, two groups viz; experimental group and control group are selected and the treatment is introduced into the experimental group only. The dependent variable is then measured in both the groups at the same time. Treatment impact is assessed by subtracting the value of the dependent variable in the control group from its value in the experimental group. Here, the two groups are identical with respect to their behaviour towards the phenomenon considered. The advantage of this design is that the data can be collected without the introduction of problems with the passage of time.

In *before-and-after with control design*, two groups are selected and the dependent variable is measured in both the groups for an identical time-period before the treatment. The treatment is then introduced into the experimental group only, and the dependent variable is measured in both for an identical time-period after the introduction of the treatment. The treatment effect is determined by subtracting the change in the dependent variable in the control group from the change in the dependent variable in experimental group. This design avoids extraneous variation resulting both from the passage of time and from non-comparability of the experimental and control groups.

(b) Formal Designs

The ***completely randomized design*** (CRD) is the simplest of all the experimental designs based on the principles of replication and randomization. Let us suppose that there are k treatments and the i^{th}

treatment is being replicated n_i times. In CRD, the whole of the experimental units are distributed completely at random to the treatments subject to the condition that the i^{th} treatment occurs n_i times. Randomization assures that extraneous factors do not continually influence one treatment. Equal number of replications for each treatment should be made except in particular cases when some treatments are of greater interest than the others. The layout of the data in CRD is that shown in the demonstration of ANOVA technique. The analysis of the data is analogous to the ANOVA for one-way classification. If some of the observations for any treatment are lost, then the standard analysis can be carried out on the available data. If the experimental material is not homogeneous, this design suffers from the disadvantages of being inherently less informative. Randomization is not restricted in any direction to ensure that the units receiving one treatment are similar to those receiving the other treatments.

If the whole of the experimental material is not homogeneous, then the ***randomized block design*** (RBD) is the simple method of controlling the variability of the experimental material. The RBD consists in grouping the experimental material into relatively homogeneous strata called blocks and the treatments are applied at random to the units within each block and replicated over all the blocks. The treatments are allocated at random within the units of each block and thus randomization is restricted. Also, variation among blocks is removed from error variation. Hence, if it is desired to control one source of variation by stratification, the experimenter should select RBD rather than CRD. The *layout* of RBD is as given below:

	Treatments						
Blocks	1	2	-	-	-	k	Total
1	x_{11}	x_{21}	-	-	-	x_{k1}	$T._1$
2	x_{12}	x_{22}	-	-	-	x_{k2}	$T._2$
-	-	-	-	-	-	-	-
-	-	-	-	-	-	-	-
r	x_{1r}	x_{2r}	-	-	-	x_{kr}	$T._r$
Total	$T_{1.}$	$T_{2.}$	-	-	-	$T_{k.}$	T

The *statistical model* assuming the various effects to be additive is given by,

$$x_{ij} = \mu + t_i + b_j + \in_{ij}$$

Where x_{ij} is the observation of the j^{th} unit getting the i^{th} treatment, μ is the general mean effect, t_i is the effect due to the i^{th} treatment, b_j is the *effect due to the* j^{th} block and $\in_{ij}$ is the random effect on the x_{ij}^{th} unit.

The *statistical analysis* of the data in RBD with one observation per experimental unit is analogous to the ANOVA for two-way classification data as shown below.

Sources	df	SS	MSS	F-ratio
Treatments	(k-1)	SS_{Tr}	MSS_{Tr}	F_{Tr}
Blocks	(r-1)	SS_{Bl}	MSS_{Bl}	F_{Bl}
Error	(k-1)(r-1)	SS_E	MSS_E	–
Total	(kr-1)	SS_T	–	–

where,

$$SS_T = \Sigma x^2_{ij} - \frac{T^2}{kr}$$

$$SS_{Tr} = \Sigma \frac{T^2_{i.}}{r} - \frac{T^2}{kr}$$

$$SS_{Bl} = \Sigma \frac{T^2_{.j}}{k} - \frac{T^2}{kr}$$

and $$SS_E = SS_T - SS_{Tr} - SS_{Bl}$$

The chief advantages of RBD are accuracy and ease of analysis. The RBD is not suitable for large number of treatments or for cases in which complete block contains considerable variability.

(c) Factorial Experiments

In factorial experiments, the effects of several factors of variation are studied and investigated simultaneously. Here, the *treatments are the combinations* of different factors under study. In these experiments, an attempt is made to estimate the effects of each of the factors and also their *interaction effects*. Let us suppose that there are p different doses of diazepam and q different doses of nitrozepam. The p and q are termed as the levels of the factors diazepam and nitrozepam respectively. A series of experiments in which only one factor is varied at a time would be lengthy, costly and unsatisfactory because of systematic change in the general background conditions. Moreover, these simple experiments do not tell us anything about the interaction effect. Alternatively, we try to investigate the variations in several factors simultaneously by conducting a $p \times q$ *factorial experiment*. In general, if the levels of various factors are equal, then r^s experiment means a factorial experiment with s factors each at r levels.

Let us deal with 2^2 *factorial experiment*. In this design, we have two factors each at two levels, and hence there are four treatment

combinations in all. Let the capital letters A and B indicate the names of the two factors under study. Let the small letters a and b denote one of the two levels of each of the corresponding factors and this will be usually called the second level. The first level of A and B are generally expressed by the absence of the corresponding small letters in the treatment combinations. The four treatment combinations may be enumerated as follows:

1 : Both factors A and B are at first level

a : A is at second level and B is at first level

b : A is at first level and B is at second level

ab : Both A and B are at second level

The four treatment combinations can be compared by laying out the experiment in CRD or RBD with r replications and ANOVA can be carried out accordingly. In factorial experiments, our main objective is to carry out separate tests for the main effects A, B and the interaction AB by splitting the SS_{Tr} with 3 df into three orthogonal components each associated either with the main effects A,B or the interaction AB, and each with one degree of freedom. Let us deal with the experiment in RBD. The layout is as given below.

	Treatments				
Blocks	1	a	b	ab	Total
1	x_{11}	x_{21}	x_{31}	x_{41}	$T._1$
2	x_{12}	x_{22}	x_{32}	x_{42}	$T._2$
–	–	–	–	–	–
–	–	–	–	–	–
r	x_{1r}	x_{2r}	x_{3r}	x_{4r}	$T._r$
Total	(1)	(a)	(b)	(ab)	T

The analysis of the data in 2^2 experiment is as given in the following ANOVA table

Source	df	SS	MSS	F-ratio
Treatments:	3	SS_{Tr}	MSS_{Tr}	F_{Tr}
Main A	1	SS_A	MSS_A	F_A
Main B	1	SS_B	MSS_B	F_B
Interaction AB	1	SS_{AB}	MSS_{AB}	F_{AB}
Blocks	(r-1)	SS_{Bl}	MSS_{Bl}	F_{Bl}
Error	3(r-1)	SS_E	MSS_E	–
Total	(4r-1)	SS_T	–	–

where,

$$SS_A = [-(1) + (a) - (b) + (ab)]^2/4r$$

$$SS_B = [-(1) - (a) + (b) + (ab)]^2/4r$$

$$SS_{AB} = [(1) - (a) - (b) + (ab)]^2/4r$$

(d) Clinical Trials

The clinical trial is a carefully and ethically *designed experiment* with the aim of answering some precisely framed question. They include both the therapeutic and prophylactic trials.

(i) Therapeutic Trials

The therapeutic trials are the *tests of remedies* on a disease or a disorder that is followed whenever some improved technique is reported. Such tests of treatment methods differ little from the regular clinical practice. The differences seem largely a matter of degree of attention to details of whom to test, how to treat and of how to measure the effect of treatment.

A *control therapy* is essential in controlled trials. If no satisfactory placebo exists, them the new treatment should be compared with dummies to check whether it has any appreciable effect on the disease. When a treatment is already in existence, the new form of treatment should normally be compared with the established one. As a first step, the *eligible patients* from the available patients are selected as per the medical and ethical considerations. Then a list of *volunteers* must be listed. In order to overcome the ethical problems, the volunteers must be informed that they may be assigned to either the experimental or the control group and a written consent of the subject to participate must be obtained. The *volunteers are stratified* according to the objectives of the trial. In each group, allot the treatments and controls at random in such a way that the groups are initially equivalent in all respects relevant to the enquiry. In order to reduce the bias, the experimenter may prefer *blind trials*. In single blind study, only the investigator is aware of what is administered to each subject. In double blind study, neither the subjects nor the investigator know the identity of what is administered.

The *mode of administration* of therapy is an important aspect in any therapeutic trial. In case of new treatment, the information is at first scanty. The testing may reveal some of the dangers of side-effects of the drug. We may choose one dose of a drug out of many, vary the interval of its administration, given it by different routes for different lengths of time and so on. Identical procedures should be followed for the control group also. In *observing the findings*, standard record forms must be drawn up and uniformity in completing must be maintained in both the groups. Every departure from the design of the experiment lowers its efficiency to some extent. The deviations from the procedures laid down may be due to the removal of the patients due to serious side effects, deaths, deterioration of conditions etc. These deviations must be considered in the *statistical analysis* of data and valid inferences have to be made.

(ii) Prophylactic Trials

The prophylactic trials attempt to *prevent* some diseases from occurring by applying a treatment to susceptible persons before that disease has appeared. These trials in medical practice are most convenient to run when the case load of patients at risk is high, the risk factor is not low and the time between cause and effect is short. The basic principles of these trials are the same as that of the therapeutic trials. In prophylactic trials, the susceptible people are divided into two groups and one group is given the prophylactic and the other is not (control). These groups are then observed over the same period to see if the protected group has suffered a lower incidence of the particular disease than the unprotected. These trials have been carried out particularly in infectious diseases.

CHAPTER 16

OBSERVATIONAL STUDIES

In observational studies, the investigator observes the events as they occur and hence the *population is not formed purposefully* in a specified way. These studies have vital role to play in identification of causal factors of mental/behavioural disorders and determination of predictors of outcome. The correlations which are observed in such studies may not directly indicate cause-and-effect phenomenon. It will be necessary to examine and try to elucidate the multiplicity of factors influencing the results. Depending on the time frame, these studies may be classified as cross-sectional, retrospective and prospective studies.

(a) Cross-sectional Studies

In cross-sectional studies, the population under study is observed, examined, investigated or questioned in order to obtain information about the population with reference to the *time of the survey*. For example, a survey carried out on in-patients of government mental hospitals in India with regard to the frequency distribution of length of stay of these patients at that time. In such a survey carried out on first july 1999, there were 15,345 in-patients in all the government mental hospitals (N=36) in India. Out of these, 52% were for below 2 years (non-chronic

patients), 18% were in two to five years duration of stay, another 18% were in five to fifteen years duration of stay and the remaining 12% were staying for more than fifteen years. The morbidity studies are cross sectional in nature. These studies may be employed to determine association between different diseases but they cannot determine which disease might have occurred first. They provide limited information to determine etiological factors. These studies are cost effective and provide economy of time and labour.

(b) Retrospective Studies

In retrospective surveys (case-control studies), the population under study is observed, examined, investigated or questioned in order to obtain information about the population with reference to the characteristics which occurred *previous to the time of the survey.* These studies attempt to compare the frequency/quantity of the suspected cause in the affected group with that of the non-affected group in the population from which the patients come. These studies are symbolically presented as shown below.

Past		Present	
Cause ?	←	Effect	Experimental group
Cause ?	←	No effect	Control group

In these studies, an estimation of relative risks known as *odds ratio* is calculated. The odds ratio is denoted by ω . It is given by,

$$\omega = \frac{\text{Odds that exposed individuals will have disease}}{\text{Odds that non-exposed individuals will have disease}} = \frac{n_{11} \times n_{22}}{n_{12} \times n_{21}}$$

where the data format is presented in 2×2 table as shown below.

	Cause (retrospectively noticed)	
Effect	Present	Absent
Present	n_{11}	n_{12}
Absent	n_{21}	n_{22}

The ω lies between zero and infinity. In *testing the significance of odds ratio*, the null and the alternate hypotheses are specified as,

$$H_0 : \omega = 1 \text{ against } H_1 : \omega \neq 1$$

We have the transformation,

$$\lambda = \ln \omega$$

Now the null and the alternate hypotheses are specified as,

$$H_0 : \lambda = 0 \text{ against } H_1 : \lambda \neq 0$$

The critical ratio and its distribution is given by,

$$Z = \frac{\hat{\lambda}}{\sqrt{\text{Var}(\hat{\lambda})}}$$

Where $\hat{\lambda}$ is based on the sample values and

$$\text{Var}(\hat{\lambda}) = \left[\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}} \right]$$

Let us make an hypothesis that puerpheral psychoses is associated with marital disharmony. There were 2465 married female registrations during the year 1975 at NIMHANS hospital. The details with respect to the occurrence of puerpheral psychoses and marital disharmony are given in the following table.

Puerpheral psychoses	Marital disharmony Present	Absent	Total
Present	8	26	34
Absent	49	2382	2431
Total	57	2408	2465

For this data, the odds ratio is calculated as,

$$\omega = \frac{8 \times 2382}{26 \times 49} = 14.956$$

That is, marital disharmony patients had the risk of puerpheral psychoses about 15 times more than those of marital harmony patients. In testing the significance of this ratio, we have

$$\hat{\lambda} = \ln 14.956 = 2.705$$

$$\text{Var}(\hat{\lambda}) = \left[\frac{1}{8} + \frac{1}{26} + \frac{1}{49} + \frac{1}{2382}\right] = 0.184$$

Thus, $$Z = \frac{2.705}{\sqrt{0.184}} = 6.301$$

Since the observed z value is more than 2.58, the null hypothesis is rejected. Hence, the alternative hypothesis 'marital disharmony and puerpheral psychoses are associated' is accepted with more than 99% confidence ($P < 0.01$).

The *determination of the population* from which the patients come is a crucial step in the design of a retrospective study. Also the choice of a suitable control group as a representative sample of the above population is also very important. Any mistakes committed at this stage will introduce bias into the investigation, resulting in artificial relationship.

The control group may be obtained by random sampling, by using matching techniques or patients with other comparable diseases. The results of a retrospective study may point to the existence of a real association between a disease and some etiological factor, but they can *never prove* cause and effects. These studies are cost effective, and economy of time and labour.

(c) Prospective Studies

The prospective surveys (cohort studies, incidence studies, longitudinal or follow-up studies) start with a defined population and the *population is followed up* in time and observed at specified times or continuously according to the design of the survey. The changes undergoing in the population are noted down according to a predetermined schedule. For example, surveys carried out successively at each delivery to detect females found to be suffering from puerpheral psychoses, out of an initial population found to have been marital disharmony at a given time and similar surveys carried out in a known population of married females with marital harmony to use as a control. Symbolically, these surveys are presented as follows.

	Present		Future
Experimental group	Cause	→	Effect ?
Control group	No cause	→	Effect ?

When the data are in the form of counts, the strength of the correlation between the risks factor and the disease may be expressed as the *relative risk* (RR) defined as,

$$RR = \frac{\text{Incidence of disease among exposed}}{\text{Incidence of disease among non-exposed}} = \frac{n_{11} \times n_{22}}{n_{12} \times n_{21}}$$

where the data format is given below.

Cause	Effect (prospectively noticed)	
	Present	Absent
Present	n_{11}	n_{12}
Absent	n_{21}	n_{22}

The cohort design permits absolute as well as relative comparisons of disease *incidence* among the exposed and the unexposed. These studies have great value when a precise hypothesis has already been formulated and it is desirable to obtain clear evidence to support or refute it. When the disease is rare, a prospective study may not be possible because the required minimum number of cases will not be available from the population. These designs are not suitable for a pilot study. These surveys are costly, time consuming and labour involved, but advantages of complete and accurate history provision.

CHAPTER 17

MULTIVARIATE STATISTICAL METHODS

All statistical methods which simultaneously analyze *more than two variables* are categorized as multivariate statistical methods. These methods take into account of the various correlations among variables. A series of univariate analysis in which each variable is analyzed at a time may give wrong interpretations of the results. This is because univariate analyses do not consider the *correlations among the variables.* The multivariate statistical methods are empirical and deal with reality. They give realistic results. Besides being a tool for analyzing complex data, these methods also help in various types of decision-making. The basic objective underlying these methods is to represent a collection of massive data in a simplified way. Their applications have been *accelerated* in modern times because of the advent of high-speed electronic computers. Most of these methods are the generalizations of univariate / bivariate statistical methods.

(a) Profiles

The profile technique simultaneously *plots data on several variables.* Thus a profile is best described as histogram on each variable, connecting between variables by identifying cases. The various objectives of profile techniques make it difficult to give complete explicit instructions for constructing profiles. It is especially useful, in clustering, in informally

suggesting possible clusters of similar cases and also clusters of similar variables. The percentage frequencies on gender (male), income per month (more than one thousand rupees), and type of service (in-patients) of five diagnostic blocks, viz: organic psychoses (OP), substance use disorders (SUD), schizophrenic disorders (Sch), affective disorders (AD), and neurotic disorders (ND) of 409 psychiatric patients registered at NIMHANS hospital are shown in the following table.

(Figures in %)

Characteristics	OP (14)	SUD (58)	Sch (124)	AD (139)	ND (74)
1. Males	79	98	53	53	45
2. Income per month (above Rs.1000)	21	31	36	44	53
3. In-patients	43	71	40	30	11

The initial analysis consists of single-variable summaries such as minimum and maximum percentages. The profile diagram is drawn as shown in Figure 17.1.

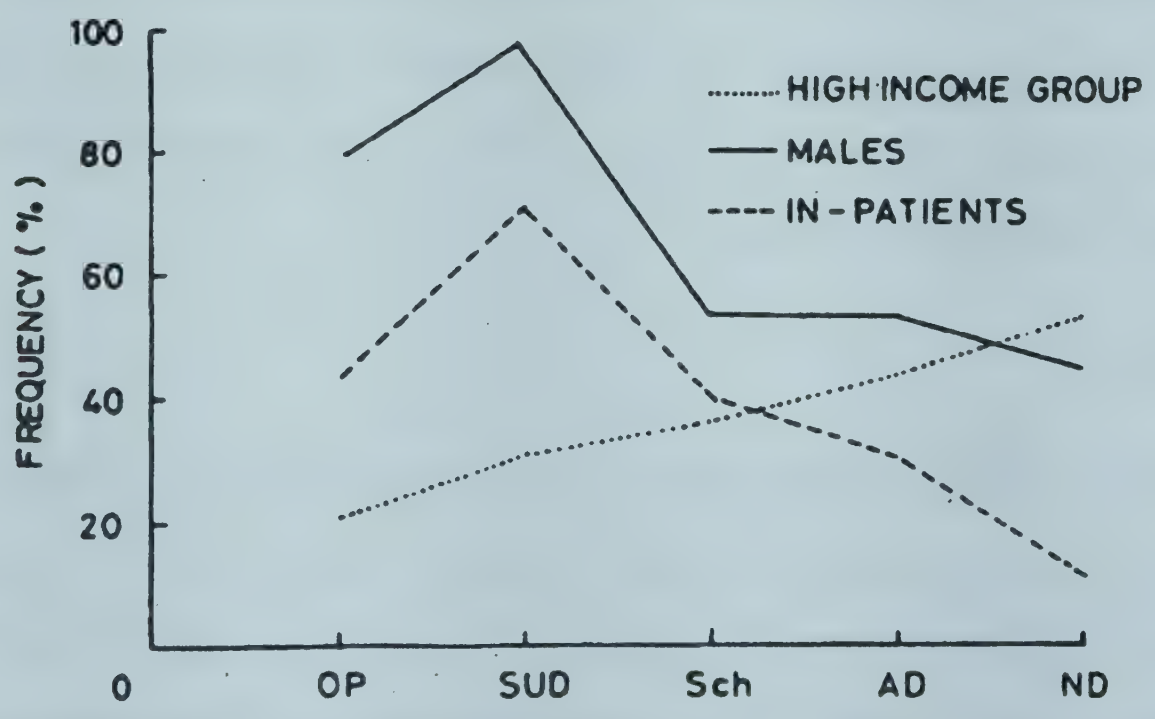


Figure 17.1 Profile Showing the Percentage Frequencies of Five Diagnostic Blocks on 3 Variables of Psychiatric Patients of NIMHANS Hospital.

It can be grasped that the variables male gender and in-patients are similar. Further, it may be grasped that the schizophrenic disorders and affective disorders are similar with respect to these three variables.

(b) Partial and Multiple Correlation Coefficients

In *trivariate data*, the correlation coefficient between x_1 and x_2 after eliminating the effect of x_3 is called partial correlation coefficient denoted by $r_{12.3}$. It is given by

$$r_{12.3} = \frac{r_{12} - r_{13}\, r_{23}}{\sqrt{(1 - r_{13}^2)\,(1 - r_{23}^2)}}$$

The correlation coefficient between x_1 and the combined influence of x_2 and x_3 is called multiple correlation coefficient denoted by $R_{1(23)}$. It is given by,

$$R^2_{1(23)} = \frac{r_{12}^2 + r_{13}^2 - 2\, r_{12}\, r_{13}\, r_{23}}{(1 - r_{23}^2)}$$

Let us suppose that x_1, x_2 and x_3 be the heights of father, mother and son respectively. With usual notations, their correlations are given by $r_{12} = 0.15$, $r_{13} = 0.60$ and $r_{23} = 0.45$. Then the partial correlation coefficient of heights between father and son after eliminating the influence of mother is calculated to be,

$$r_{13.2} = \frac{0.60 - 0.15 \times 0.45}{\sqrt{(1 - 0.15^2)\,(1 - 0.45^2)}} = 0.603$$

The correlation coefficient of heights between son and the combined influence of father and mother is calculated to be,

$$R^2_{3(12)} = \frac{0.60^2 + 0.45^2 - 2 \times 0.15 \times 0.60 \times 0.45}{(1 - 0.15^2)} = 0.493$$

Hence, $R_{3(12)} = \sqrt{0.493} = 0.702$

(c) Multiple Regression Analysis

The multiple regression analysis is based on the assumption that all the variables are normally distributed. A regression of k independent variables ($x_1, x_2, \ldots x_k$) upon a dependent variable y may be expressed as,

$$y = a + b_1 x_1 + b_2 x_2 + \ldots\ldots + b_k x_k + e$$

Where a is a constant term and b_j are called the sample partial regression coefficients. The normal equations which yield a least square solution are given by (in matrix form),

$$\begin{bmatrix} a \\ b_1 \\ - \\ - \\ - \\ b_k \end{bmatrix} = \begin{bmatrix} n & \Sigma x_1 & - & - & - & \Sigma x_k \\ \Sigma x_1 & \Sigma x_1^2 & - & - & - & \Sigma x_1 x_k \\ - & - & - & - & - & - \\ - & - & - & - & - & - \\ - & - & - & - & - & - \\ \Sigma x_k & \Sigma x_k x_1 & - & - & - & \Sigma x_k^2 \end{bmatrix}^{-1} \begin{bmatrix} \Sigma y \\ \Sigma y x_1 \\ - \\ - \\ - \\ \Sigma y x_k \end{bmatrix}$$

Let us *suppose* that a psychiatrist wishes to determine the type of patients who respond well to his treatment program. He administers an intelligence quotient test (x_1) and a personality test (x_2) to a group of five patients. At the end of the program he rates the success of the treatment (y) on a 7-point scale. The scores are as given in the following table .

Patients	IQ (x_1)	Personality (x_2)	Success (y)
A	91	4	7
B	100	2	4
C	109	3	1
D	97	0	3
E	103	6	5
Mean	100	3	4
SD	6	2	2

We solve $b = R^{-1} k$

Where R is the correlation matrix relating to independent variables and k is the vector that contains the correlation between the independent variables and the dependent variable.

In this example

$$b = \begin{bmatrix} 1.00 & 0.15 \\ 0.15 & 1.00 \end{bmatrix}^{-1} \begin{bmatrix} -0.80 \\ 0.45 \end{bmatrix} = \begin{bmatrix} -0.89 \\ 0.58 \end{bmatrix}$$

The elements of b are the regression weights that we apply to a persons standard scores in assigning the probable outcome of his treatment in the standard form. The standard scores of the five patients are calculated as shown in the following table.

Patients	IQ (X_1)	Personality (X_2)	Success (Y)
A	–1.5	0.5	1.5
B	0	–0.5	0
C	1.5	0	–1.5
D	–0.5	–1.5	–0.5
E	0.5	1.5	0.5

The observed scores (Y) and predicted scores ($\hat{Y}$) of success in the standard form of the five patients are calculated as shown in the following table.

Patients	Y	$\hat{Y}$	$(Y - \hat{Y})$
A	1.5	1.625	–0.125
B	0	–0.290	0.290
C	–1.5	–1.335	–0.165
D	–0.5	–0.425	–0.075
E	0.5	0.425	0.075

The predicted scores ($\hat{y}$) of success of the five patients are obtained by multiplying $\hat{Y}$ by the standard deviation and adding the mean of the dependant variable as shown in the following table.

Patients	y	$\hat{y}$	$y - \hat{y}$
A	7	7.25	-0.25
B	4	3.42	0.58
C	1	1.33	-0.33
D	3	3.15	-0.15
E	5	4.85	0.15
Total	20	20.00	0.00

The estimates are fairly close to the actual values and the correlation between the two is 0.986, which is the multiple correlation coefficient denoted by R. This is the maximum possible correlation between the dependant variable and any linear combination of the independent variables. In practice, R is calculated by finding the inner product of vector k and vector b, and taking the square root of this product. That is ,

K	b	Product
–0.80	–0.89	0.712
0.45	0.58	0.261

$R^2 = 0.973 \quad R = 0.986$

The values of 0.712 and 0.261 are known as coefficient of separate determination. It is some times useful to express the contribution of a variable to the prediction by expressing its coefficient of separate determination as a percentage of R^2. Thus,

$$\text{Contribution of IQ} = \frac{0.712 \times 100}{0.973} = 73\ \%$$

Some times this method of assessing the relative contribution of variable can lead to apparently absurd results when any of the coefficients of separate determination is a negative value.

(d) Multivariate Analysis of Variance

The means of a sample on several variables may be thought of as a single point in a space that has as many dimensions as there are variables. The multivariate analysis of variance (MANOVA) technique may be employed to test the significance of the difference between *multivariate means of several groups*. In order to perform this analysis, we first calculate the pooled within group sum of squares and sum of products matrix W. Then we calculate an L-criterion as the ratio of within to total sums of squares and cross products. That is ,

$$L = \frac{|W|}{|T|}$$

L lies between 0 and 1. If L is one, then there is no between groups variances or covariances. This means that every group has the same mean score on a particular variable and that this is true for all variables. In order to make an approximate test of the null hypothesis, we calculate,

$$\chi^2 = -(\sum n_g) \ln L \qquad \text{with } df = (g-1)t$$

The scores of three art graduates and three science graduates on two aptitude tests x and y are given below:

Arts graduates	Scores	
	x	y
P_1	3	6
P_2	4	4
P_3	5	8

Science graduates	Scores	
	x	y
P_4	1	2
P_5	2	0
P_6	3	4

$$W = W_1 + W_2 = \begin{bmatrix} 2 & 2 \\ 2 & 8 \end{bmatrix} + \begin{bmatrix} 2 & 2 \\ 2 & 8 \end{bmatrix} = \begin{bmatrix} 4 & 4 \\ 4 & 16 \end{bmatrix}$$

$$T = T_1 + T_2 = \begin{bmatrix} 5 & 8 \\ 8 & 20 \end{bmatrix} + \begin{bmatrix} 5 & 8 \\ 8 & 20 \end{bmatrix} = \begin{bmatrix} 10 & 16 \\ 16 & 40 \end{bmatrix}$$

$$L = \frac{|W|}{|T|} = \frac{(64 - 16)}{(400 - 256)} = 48/144 = 1/3$$

$$\chi^2 = -6 \ln (1/3) = -6 (-1.0986) = 6.592$$

with df = (2–1) 2 = 2

Since $P < 0.05$, we have the evidence that the two groups are drawn from populations which differ in their mean quantities of x or in their mean quantity of y or in some compound of the two.

CHAPTER 18

CLUSTER ANALYSIS

Cluster Analysis encompasses many diverse techniques for *discovering structure* within the complex bodies of data. In a typical situation, one has a sample of data units (for example, psychiatric patients) each described by scores on selected variables such as achievements test scores. The *objective* is to group either the data units or the variables into clusters such that the elements within a cluster have a high degree of natural association among themselves while the clusters are relatively distinct from one another. Thus the clustering process generates a new categorical scheme or recovers groups within a mixture of several populations.

The term *classification* refers to an arrangement of classes already known or sorting individual objects into well-known classes. *Identification* is the allocation of individual objects to established classes on the basis of specific criteria. In psychiatry, the same process is called as diagnosis, and it refers to the identification of a familiar disorders from the symptoms presented by the patient. The *discriminate analysis* deals with the determination of a set of characteristics that can significantly differentiate groups. The cluster analysis methods are also used to group similar variables to define dimensions or factors. Hence they are also called as poor man's *factor analysis.*

The cluster analysis may be used for a variety of *research goals*. The best known of these research goals is the making of *classifications*. Researchers in all fields need to make and revise classifications continuously. For example, psychiatric researchers need to make classification of mental patients based on psychiatric test scores in order to improve their understanding of mental illness and plan for its treatments. The proportional incidence of one symptom or another indicates whether it is a characteristic feature of the disease or not. The identification of signs and symptoms of a disorder or syndrome is the major concern of psychiatrists. Inadequate parental control of hyperactivity children is found by chance and it is found in almost every case of conduct disorder children.

The cluster analysis may be useful in *shedding light* on the previously made hypotheses. For example, there has long been controversy over the classification of depressed patients. The issues involved here have been reviewed on a number of occasions. Several attempts have been made to establish the validity of classifying such patients into reactive and endogenous groups. The problem has been tackled with some success by using cluster analysis techniques.

The cluster analysis may be useful to produce groups which form the basis of a classification scheme useful in latter studies for *predictive purposes* of some kind. For example, a cluster analysis applied to data consisting of a sample of psychiatric patients may produce groups of patients which react differently when treated with some drug. Thus it enables us to decide whether a drug is suitable for a particular type of patients. Such a procedure was used in an investigation of the usefulness of amytriptaline in the treatment of depression.

The researcher may use cluster analysis to *explore the data and generate hypotheses, to fit models, to summarize and display the data, and to name the relevant groups and give explanations.*

There are mainly two *clustering structures* called the tree and the partition. A cluster is a subset of a set of objects. A *tree* is a family of clusters which has the property that any two clusters are either disjoint or one includes the other. The hierarchical clustering techniques are used to obtain trees. A *partition* is a family of clusters which has the property that any item just belongs to one of the partitions. Thus, a partition with a set of all items added is a tree.

(a) Hierarchical Methods

In hierarchical methods, the classes themselves are classified into groups and the process is repeated at different levels to form a tree. There are two types of hierarchical methods, viz; the *agglomerative techniques and the divisive techniques.* The agglomerative techniques proceed by a series of successive fusions of n entities into groups, and the divisive techniques partition the set of entities successively into finer partitions. Thus, the agglomerative techniques ultimately reduce the data to a single cluster containing all the entities, and the divisive techniques will finally split the entire set of data into groups each containing a single entity. The results of both agglomerative and divisive techniques may be presented in the form of a *dendrogram.* A dendrogram is a two dimensional diagram illustrating the fusions or partitions which have been made at each successive level. The basic procedure with all agglomerative methods is similar. They begin with the computation of a matrix of proximity between the entities. At any particular stage, the techniques fuse

individuals or groups of individuals which are closest. Each fusion decreases the number of groups by one.

(i) Measures of Proximity

The hierarchical clustering techniques begin with the calculation of a matrix of similarity or distance (called proximity) between *entities.* Indeed, the clustering techniques may be thought of as attempts to summarize the information on the relationships between entities so that these relationships can be easily comprehended and communicated.

An *association coefficient* measures the similarity between the individuals, given the values of a set of p variates common to both. In many cases, the variates are of the presence and absence type which may be arranged in the familiar two way association table.

	Individual j	
Individual i	Present	Absent
Present	a	b
Absent	c	d

Many different coefficients have been suggested for data of this type. The *Simple Matching Coefficient* (SMC) and the *Jaccord's Coefficient* (JC) are the important association coefficients. They are defined as,

$$SMC = (a+d)/(a+b+c+d)$$

$$JC = a/(a+b+c)$$

An association coefficient usually takes values in the range zero to one. If simple matching coefficient is used, some cases would appear very similar primarily because they both locked the same features rather than because of the features they did have well shared. In contrast, JC is

concerned only with features that have positive co-occurrences.

The *distance measures* are the measures of dissimilarities. Normally, they have no upper bounds and they are scale dependent. The *Euclidian Distance* (ED) is the commonly used measure of distance. The Euclidian Distance (ED_{ij}) between two individuals i and j is given by,

$$ED_{ij} = \sum_{k}(x_{ik} - x_{jk})^2$$

Where x_{ik} is the value of the k^{th} variable for the i^{th} individual. The Euclidian Distance used on raw data may be very unsatisfactory since it is badly affected by the change of scale of a variable. Because of this, variables are frequently standardized to have zero mean and unit variance. Although this has problems, the Euclidian Distance calculated from the standardized variables will preserve relative distances. The *Absolute Distance* (AD_{ij}) between two individuals i and j is given by,

$$AD_{ij} = \sum_{k}| x_{ik} - x_{jk}|$$

The *choice of measure of proximity* is an important consideration since different measures may lead to different results on the same data. It may be proposed that both the association coefficients and Euclidian Distance is used as proximity of profile data so that it is possible to determine the correct measure to be used in future. The *choice of variables* used to cluster individuals plays a crucial role in cluster analysis. The choice of the particular set of variables used to describe each entity reflects the investigators judgement of relevance for the purpose of the classification. Ideally variables should be chosen within the context of an explicitly stated theory that is used to support the classification. If

the data are not of the same scale values, they are commonly standardized. The *standardization* to a mean of zero and to unit variance can reduce the differences between groups on those variables that may well be the best discriminators of group differences. Users with substantially different units of measurement will undoubtedly want to standardize them, especially if a similarity measure such as euclidian distance is to be used.

(ii) Agglomerative Techniques

Differences between hierarchical agglomerative techniques arise because of the different ways of defining proximity between groups of individuals. In *single linkage method*, the distance between groups is defined as the distance between their closest members. This method can be used with both similarity and distance measures. This method is invariant to monotonic transformations of the proximity matrix. The major drawback of this method is that it has the tendency to claim or form long and elongated clusters. In *complete linkage method*, the distance between groups is defined as the distance between their most remote pair of individuals. Hence it is the logical opposite of the single linkage method. The major criticism of this method is that it is a space-diluting method. This method has a tendency to find relatively compact and hyperspherical clusters composed of highly similar cases. A visual examination of the Dendrogram gives a clear sense of clusters in the data and hence this method can be used to indicate the number of appropriate clusters present in the data.

In *average linkage between merged groups*, the distance between the two groups is defined as the average of the distances of all pair-wise combinations between the individuals in the two groups. Thus, here a

cluster is defined as a group of entities in which each member has a greater mean similarity with a members of the same cluster than it does with all members of any other cluster. In *average linkage within the new group,* the distance between two groups is defined as the average of the distances of all pair wise combinations of all individuals in the two groups. In practice, this method frequently gives results that are little different from those obtained with the complete linkage method.

The *minimum variance method* (Ward method) is the commonly used method in mental health care field. This method proposes that the loss of information which results from the grouping of individuals into clusters can be measured by the total sum of squares of the deviations of every point from the mean of the cluster to which it belongs. At each step in the analysis, union of every possible pairs of clusters is considered and the two clusters whose fusion results in the minimum increase in the error sum of squares are combined. Thus, this method is designed to generate clusters in such a way that the variance within a cluster is minimum. This method tends to find clusters of relatively equal size and space as hyperspheres. A visual examination of the Dendrogram from this method gives a sense of clusters. A common problem associated with the use of this method is that the clusters found by this method can be ordered in terms of their overall elevation. This method is widely used in mental health care research.

The applications of these methods in psychiatric research, have clearly shown that the *complete linkage method using the Euclidian Distance* yielded the best results. Let us demonstrate this technique with the following hypothetical data. The data consists of present (denoted by 1) and absent (denoted by -) on each of the eight symptoms for each of the five patients.

	Symptoms							
Patients	1	2	3	4	5	6	7	8
A	1	–	1	1	–	1	1	1
B	–	–	1	1	–	1	1	1
C	1	1	1	1	1	1	–	1
D	1	–	–	1	1	–	–	–
E	1	1	–	1	1	1	–	1

The *euclidian distance* matrix is as given below.

	A	B	C	D	E
A	–				
B	1	–			
C	3	4	–		
D	5	6	4	–	
E	4	5	1	3	–

As a first stage of the complete linkage procedure, individuals A and B (or C and E) are fused to form a cluster (AB) since the distance between them is the smallest. The distance between the cluster AB and the remaining three individuals C, D and E, are obtained as shown in the following table.

	AB	C	D	E
AB	–			
C	4	–		
D	6	4	–	
E	5	1	3	–

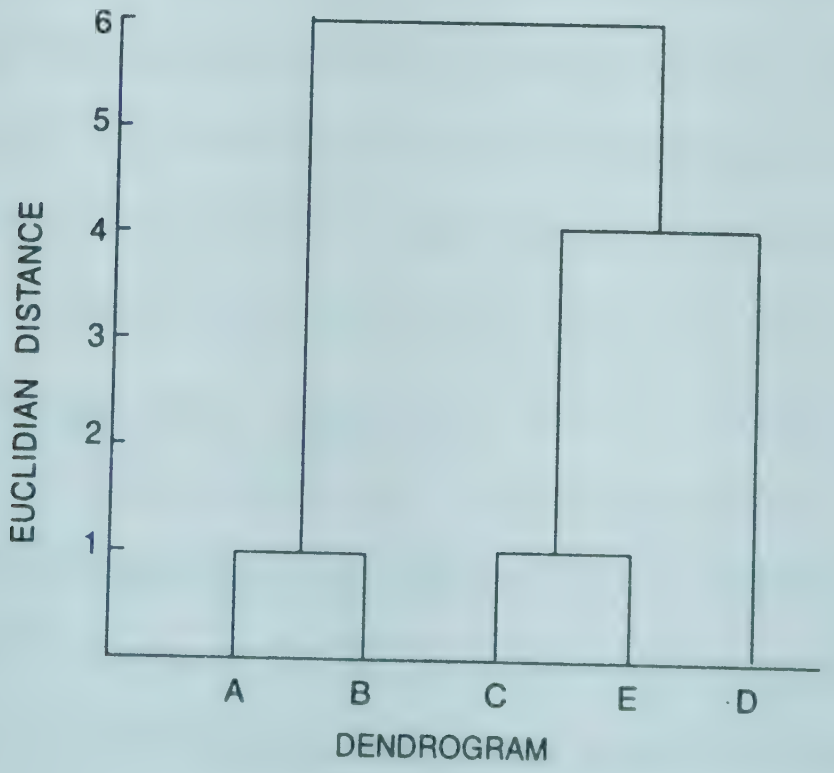
The smallest entry in the above cited matrix is 1 which is the distance between the individuals C and E. Hence, they are fused to obtain the second cluster CE. Now, the distance matrix is as given below:

	AB	D	CE
AB	–		
D	6	–	
CE	5	4	–

It can be noted that the distance between the individual D and the cluster CE is the smallest. They are fused to form the third cluster (CDE). The distance matrix is as given below:

	AB	CDE
AB	-	
CDE	6	-

Finally, the clusters AB and CDE are fused at a distance of 6 units to form a single cluster consisting of all the units. The dendrogram for the data is drawn as shown below:



The *major problem* of hierarchical methods is that there is no provision for reallocation of entities who may have been poorly classified at an early stage in the analysis. These methods can generate different solutions simply by rearranging the data in the similarity matrix. They are not stable when cases are dropped out of the analysis.

(iii) Number of Clusters

Hierarchical methods give configuration for every number of clusters from one upto the number of entities in the data. On the other hand, some other clustering methods such as the partitioning methods find a best fitting structure for a given number of clusters. We usually want the clusters to be *few in number* and to be well defined. However, the purpose of the classification is also a factor. Ultimately, we want the classification to be judged valid. Thus, we want to get 'number of clusters' that makes it to work the best.

(b) Partitioning Methods

The partitioning methods are designed to cluster data units into a single *classification of k clusters*, where k is either specified a prior or is determined as part of the hierarchical clustering method. The central idea in most of these methods is to choose some initial partition of the data units and then alter cluster membership so as to obtain a better partition. Various algorithms which have been proposed differ as to what constitutes a 'better partition' and what methods may be used for achieving improvements. The partitioning methods admit reallocation of entities and thus allowing poor initial partitions to be corrected at a later stage. Most of these methods employ distinct procedures with respect to the method of initiating clustering and a method of reallocating

some or all of the entities to other clusters once an initial classificatory process has been completed.

(i) Methods of Initiating Clusters

There are two basic ways for initiating clusters: one is by using seed points and other by selecting an appropriate starting partition. The use of *seed points approach* consists in finding k points (seed points) in the p-dimensional space, which acts as initial estimates of the cluster centre. The following methods are representative examples of how such seed points can be generated.

* Choose the first k data units in the data set
* Subjectively choose any k data units from the data set.
* Use the centroids of the k clusters of classification obtained by hierarchical methods.

When seed points are used, entities are allocated to the seed points to those whose centre they are nearest. The seed points may remain stationary throughout the assignment of the full data set. In this process, the resulting set of clusters is independent of the sequence in which data units are assigned. The estimate of the cluster center may be updated after the addition of each entity to the cluster.

(ii) Methods for Reallocating Entities

The methods for reallocating entities deal with the ways in which cases are reassigned to clusters. There are two basic types, viz: the k-means passes and the hill climbing passes. The *k-means passes* are the commonly used partitioning methods in mental health care field. The k-means

passes are also referred to as the 'nearest centroid sorting pass'. This simply involves the reassignment of cases to the cluster with the nearest centroid. The k-means passes can be either combinatorial or non-combinatorial. The former method requires the calculation of the centroids of a cluster after each changes in its membership while the later recalculates the cluster centroid only after an entire pass through the data has been completed. One other important distinction is that the k-means passes can also be either exclusive or inclusive. Exclusive methods remove the case under consideration from the parent cluster when a centroid is computed, whereas inclusive methods include them.

(iii) Forgy's Method

Among the several methods suited to the basic problem of sorting the data units into a fixed number of clusters, the method suggested by Forgy is very commonly used. This method suggests a very simple *algorithm* consisting of the following sequence of steps:

1. Begin with any desired initial configuration. Go to step 2 if beginning with a set of seed points. Go to step 3 if beginning with a partition of the data units.

2. Allocate each data units to the cluster with the nearest seed points. The seed points remain fixed for a few cycle through the entire data set.

3. ~~Complete~~ Compute new seed points as the centroids of the cluster of data units.

4. Alternate steps 2 and 3 until the process converges. That is, continue until no data unit change their cluster membership at step 2.

It is not possible to say how many repetitions of steps 2 and 3 will be enquired to achieve convergence in any particular problem. However, empirical evidence indicates that ordinarily ten or less number of repetitions will be sufficient

As an *example* of the operation of the Forgy partitioning method, let us consider the data used for the demonstration of the complete linkage hierarchical method. This set of data is to be partitioned into two clusters such that all the members of a cluster are nearer to the mean vector of that cluster than the mean vector of any other cluster. The two starting seed points may be the patients A and C. The remaining patients are examined in sequence and allocated to the cluster whose mean vector is closest. This gives the following series of results.

	Cluster 1	Cluster 2
Stage 1: Individuals	A	C
Stage 2: Individuals	AB	C
Stage 3: Individuals	AB	CD
Stage 4: Individuals	AB	CDE

Hence, at this stage we have the initial classification with,

Cluster 1: Individuals A and B

Mean Vector (0.5 – 1 1 – 1 1 1)

Cluster 2: Individuals C,D and E

Mean Vector (1 0.67 0.33 1 1 0.67 – 0.67)

Now, each patient is tested to see whether or not he is nearer the mean vector of his own cluster than that of the other cluster. It is found that

each patient is nearer to the mean vector of his own cluster. Thus the final two clusters are the cluster 1 consists of two patients A and B, and Cluster 2 consists of three patients C,D and E.

The *problems* of partitioning techniques is that sub-optional solutions are frequently found. In determining the number of clusters, it is generally suggested that a group of the criterion value like the point biserial correlation or the number of reallocations against the number of groups will indicate the correct number to consider by showing a short increase or decrease of the clustering criterion at the correct number of clusters.

(c) General Problems, Validation Techniques

In the application of cluster analytic techniques, the expert in the applied field may think that detailed knowledge is more important than mere numerical manipulations. The techniques of cluster analysis are not based on sound probability models and the results are *poorly evaluated* and unstable when evaluated. Different clustering techniques can and do generate different solutions to the same data set and hence a number of validation techniques have been developed to provide some relief for this problem. The strategy of cluster analysis is structure seeking although its operation is structure imposing and therefore the key to using cluster analysis is to know when these groups are real and not merely imposed on the data by the method.

Various intuitively reasonable procedures have been suggested for evaluating the stability and the usefulness of the set of clusters obtained. In *method of replication*, the same data set may be cluster analyzed by different methods. The solution by the majority of the methods should

be similar to say that the data is clearly structured. In *significance tests on variables used to create clusters*, the clustering solution is validated by applying ANOVA or chi-square as tests of significance of the difference between clusters on those variables used to create clusters. Since these tests are positive, regardless of whether clusters exist in the data or not, they must be performed on a controlled data set to draw valid conclusions. In *significance tests on external variables*, the obtained clusters are compared on variables of interest which were not included in the original analysis. If difference between clusters persist with respect to these variables, then there is some evidence that a useful solution has been obtained, in the sense that by stating that a particular entity belongs to a particular cluster, we convey information on variables not used to produce the clusters. In *method of marker sample*, the sample from major clinical diagnostic categories are pooled together as 'marker sample' to aid to validate the cluster solution. An adequate solution should place dissimilar subjects into different clusters and provide some indication of the number of clusters required to provide reasonable differentiation. For example, it would have been conceptually inconsistent for a given cluster solution to classify numerous psychosomatic and psychotic patients together in the same cluster.

CHAPTER 19

DISCRIMINANT FUNCTION ANALYSIS

The discriminant function is introduced as a statistical technique to facilitate the *classification* of persons. First, let us deal with the case of two groups. Here we are given with two classes of persons and t measurements have been made on each person. We wish to find out the weighing vector which, when applied to some newly observed and unclassified person, will assign him to one or other of the classes with the smallest probability of error. We assume that, in the populations from which the classes are drawn, the t variables have a common multivariate normal distribution. The vector of weights (w) which provides the optimum assignment is given by ,

$$w = V^{-1}d$$

Where V is the weighted average of the variance-covariance matrices of the two classes and d is the vector of differences between the t pairs of means of the two classes.

Let us *suppose* that a psychiatrist has to take a decision whether conduct disorder diagnosis or hyperkinesis diagnosis may be most suitable for a child at a child guidance clinic. The information on which the decision is based is derived from the scores on two dimensions, viz: the psychosocial stresses and the helpfulness in a checklist for assessment of

childhood psychopathology at the time of registration. We take the scores of three conduct disorder children and three hyperkinesis children as the basis for our advice. The children obtain the following scores on the two dimensions of the checklist, x and y.

Conduct disorder children	x	y
C_1	3	6
C_2	4	4
C_3	5	8
Mean	4	6

Hyperkinetic children	x	y
C_4	1	2
C_5	2	0
C_6	3	4
Means	2	2

We calculate the sum of squares and sums of products matrix of each diagnostic group:

$$W_1 = \begin{bmatrix} 2 & 2 \\ 2 & 8 \end{bmatrix} \qquad W_2 = \begin{bmatrix} 2 & 2 \\ 2 & 8 \end{bmatrix} \qquad W = \begin{bmatrix} 4 & 4 \\ 4 & 16 \end{bmatrix}$$

The weighted average of the variance-covariance matrix is given by,

$$V = \frac{W}{(n_1-1) + (n_2-1)} = \frac{W}{4} = \begin{bmatrix} 1 & 1 \\ 1 & 4 \end{bmatrix}$$

$$V^{-1} = \begin{bmatrix} \frac{4}{3} & -\frac{1}{3} \\ -\frac{1}{3} & \frac{1}{3} \end{bmatrix}$$

Hence
$$w = \begin{bmatrix} \frac{4}{3} & -\frac{1}{3} \\ -\frac{1}{3} & \frac{1}{3} \end{bmatrix} \begin{bmatrix} 2 \\ 4 \end{bmatrix} = \begin{bmatrix} \frac{4}{3} \\ \frac{2}{3} \end{bmatrix}$$

The discriminant function scores of the six children are calculated as shown in the following table. For example, the first child in the conduct disorder group is given by, $(4/3) \times 3 + (2/3) \times 6 = 8$

Conduct disorder children	Score
C_1	8.00
C_2	8.00
C_3	12.00
Mean	9.33

Hyperkinetic children	Score
C_4	2.67
C_5	2.67
C_6	6.67
Mean	4.00

Generally, for classification purpose, the appropriate cutting score is the half-way mark between the means of the scores of the two diagnostic groups. In our example, the half-way mark is 6.67. A good way to summarize a discriminant analysis is to draw up a contingency table. The rows of the table indicate the actual diagnosis to which a child belongs and the columns indicate the diagnosis to which the model assigns him. The contingency table for the five children in our example is as given below. Note that one hyperkinesis child has the score equal to the half-way mark score and hence cannot be discriminated.

		Predicted diagnosis	
		Conduct disorder	Hyperkinesis
Actual Diagnosis	Conduct disorder	3	0
	Hyperkinesis	0	2

(a) Mahalanobis D^2

The Mahalanobis D^2 is the inner product of the vector of difference d and the vector of weights w. That is,

$$D^2 = d' w = d' V^{-1} d$$

In our example,

$$D^2 = [2 \quad 4] \begin{bmatrix} \frac{4}{3} \\ \frac{2}{3} \end{bmatrix} = 5.33$$

Thus the D^2 is the difference between the means of the two groups on discriminant function scores. In our example, $D^2 = 9.33 - 4.00 = 5.33$

(b) Probability of Misclassification

It can be shown that *half of D* is a unit normal deviate. In our example, $0.5D = 0.5\sqrt{5.33} = 1.155$. If we look it up in a table of z- distribution (Appendix V), we find the probability that a child who really belongs to one diagnostic group will be incorrectly assigned to the other diagnostic group, by the discriminant function. In our example, the probability of misclassification is $(0.50 - 0.37) = 0.13$, which means that we may expect 87% of our assignments to be correct.

(c) Minimization of Probability of Overall Misclassification

If one group occurs *more frequently* than that of the other in the population from which we are drawing members, then the *cut-off point* should be moved towards the mean of the less frequent group. The probability of misclassifying members of the infrequent group is therefore increased, and the probability of misclassifying members of the more frequent group is decreased. By a suitable choice of cutting point, the

overall probability of misclassification may be minimized. Let us suppose that the ratio of conduct disorder children to hyperkinetic children in the population from which the random samples have been drawn is 2:1. We divide the larger of those two numbers by the smaller and take the natural logarithm. That is,

$$\ln 2 = 0.693$$

We divide this value by D. That is,

$$\frac{0.693}{2.309} = 0.300$$

We move our cutting point this distance towards the hyperkinesis disorder. Thus, the unit normal deviate for the conduct disorder diagnosis is (1.155 + 0.300) = 1.455, and the unit normal deviate for the hyperkinesis diagnosis is (1.155 – 0.300) = 0.855. The probability of wrongly classifying a conduct disorder child as a hyperkinetic child is now only (0.50 – 0.43) = 0.07, and that of the hyperkinetic diagnosis is (0.50 – 0.30)= 0.20. Since two-thirds of our population are conduct disorder children and one-third are hyperkinetic children, the estimated overall rate of misclassification, with a cut-off point $(6.67 - 0.3 \times 2.31) = 5.98$ is calculated to be $(0.07 \times 0.67) + (0.20 \times 0.33) = 0.11$ or 11%.

(d) Case of More than Two Groups

The technique of the linear discriminant function analysis developed for two groups may be generalized to the case of more than two groups. In *multiple discriminant function analysis,* a vector of weights has to be calculated for each of the groups. In order to do this, the vector d is replaced by the vector of means m_i for the i^{th} group in turn. The vector of weights for the i^{th} group is given by,

$$l_i = V^{-1} m_i$$

Let the scores of an unclassified person is given by the column vector, p. The weights of the person in the i^{th} group is given by,

$$w_i = p' l_i - \frac{1}{2} m_i' l_i$$

Then the procedure is to assign the person to the group for which he obtains the highest weighted scores.

Let us treat the problem of linear discriminant function analysis for two diagnostic groups as the problem of multiple discriminant function analysis. Let us suppose that the scores of an undiagnosed child are 2 and 5 on the two dimensions of the checklist respectively. The weights of the child in the conduct disorder diagnosis is calculated as follows.

$$l_1 = V^{-1} m_1 = \begin{bmatrix} \frac{4}{3} & -\frac{1}{3} \\ -\frac{1}{3} & \frac{1}{3} \end{bmatrix} \begin{bmatrix} 4 \\ 6 \end{bmatrix} = \begin{bmatrix} \frac{10}{3} \\ \frac{2}{3} \end{bmatrix}$$

$$w_1 = [2 \quad 4] \begin{bmatrix} \frac{10}{3} \\ \frac{2}{3} \end{bmatrix} - 0.5 \; [4 \quad 6] \begin{bmatrix} \frac{10}{3} \\ \frac{2}{3} \end{bmatrix}$$

$$= (30/3 - 26/3) = 4/3 = 1.333$$

The weights of the child in the hyperkinesis diagnosis is calculated as follows:

$$l_2 = V^{-1} m_2 \begin{bmatrix} \frac{4}{3} & -\frac{1}{3} \\ -\frac{1}{3} & \frac{1}{3} \end{bmatrix} \begin{bmatrix} 2 \\ 2 \end{bmatrix} = \begin{bmatrix} 2 \\ 0 \end{bmatrix}$$

$$w_2 = [2 \quad 2] \begin{bmatrix} 2 \\ 0 \end{bmatrix} - 0.5 \; [2 \quad 2] \begin{bmatrix} 2 \\ 0 \end{bmatrix} = 4 - 2 = 2$$

Hence, the given child in most appropriated diagnosed as hyperkinetic.

CHAPTER 20

FACTOR ANALYSIS

The factor analysis is applicable when there is a systematic *interdependence* among a set of observed variables and the researcher is interested in finding out something more fundamental which creates the commonality. This technique seeks to resolve a large number of measured variables in terms of relatively few categories known as factors. That is, this technique allows the researcher to *group variables into factors* and the factors so derived may be treated as new variables. The meaning and name of such new variable is subjectively determined by the researcher. Since the factors happen to be linear combinations of data, the coordinates of each observation or variable is measured to obtain what are called factor leadings. Such factor leadings represent the correlation between the particular variable and the factor.

The *mathematical basis* of factor analysis concerns a data matrix of scores of n persons on k measures. It is assumed that the scores on each measure are standardized. The factors are obtained by some method of factor analysis. For realistic results, we resort to the technique of rotation because such rotation reveals different structures in the data. The commonalities, the eigen values and the total sum of squares of the

loadings are obtained and the results are interpreted. Finally, factor scores are obtained which help in explaining what the factors mean. With factor scores, one can also perform several other multivariate statistical methods such as cluster analysis and multi-dimensional scaling. Factor analysis is not a single method but a set of techniques. Let us deal with the principal component analysis.

The *principal component analysis* (PCA) extracts maximum sum of squares of the loadings for each factor in turn. Accordingly, the PCA explains more variance than would the loadings obtained from any other method of factoring. It is assumed that all the variables are standardized. The aim of this method is to construct new variables (P_i), called principal components which are linear combinations of a given set of variables x_j ($j = 1,2....k$)

$$P_1 = a_{11}x_1 + a_{12}x_2 + ...+ a_{1k}x_k$$

$$P_2 = a_{21}x_1 + a_{22}x_2 + ...+ a_{2k}x_k$$

- - - - - - - - - -

$$P_k = a_{k1}x_1 + a_{k2}x_2 + ...+ a_{kk}x_k$$

The a_{ij} are called loadings and are worked out in such a way that the extracted principal components satisfy two conditions. Firstly, the PC *are uncorrelated* (orthogonal). Secondly, the first PC has the *maximum variance*, the second PC has the next maximum and so on.

Let us *suppose* that we have given two tests to each of eight persons and have obtained the following 8×2 matrix of scores, M. The scores are expressed as deviations from the test means.

$$M = \begin{bmatrix} -8 & -1 \\ 6 & 10 \\ -2 & -10 \\ 8 & 1 \\ 0 & 3 \\ -6 & -6 \\ 0 & -3 \\ 2 & 6 \end{bmatrix}$$

The sum of squares and sums of products matrix W is given by,

$$W = M' M$$

$$= \begin{bmatrix} 208 & 144 \\ 144 & 292 \end{bmatrix}$$

The total sums of squares of the two tests is 500. Within the limits set by this total, we wish to rotate each test vector in such a way that the resulting vectors (called components) satisfy two conditions. Firstly, they are orthogonal (uncorrelated). Secondly, the first component must explain the maximum variance and the second must explain the next maximum and so on. In order to do this, the matrix of scores M is multiplied by the weighing matrix F where,

$$F = \begin{matrix} f_1 & f_2 \end{matrix} \begin{bmatrix} 0.6 & 0.8 \\ 0.8 & -0.6 \end{bmatrix}$$

The component scores are given by the matrix equation C = M F. Thus,

$$C = \begin{bmatrix} -5.6 & -5.8 \\ 11.6 & -1.2 \\ -9.2 & 4.4 \\ 5.6 & 5.8 \\ 2.4 & -1.8 \\ -8.4 & -1.2 \\ -2.4 & 1.8 \\ 6.8 & -2.0 \end{bmatrix}$$

Now, $$\Lambda = C' C = \begin{bmatrix} 400 & 0 \\ 0 & 100 \end{bmatrix}$$

Thus the total sum of squares (500) is more unequally divided between the components than it was between the tests, and the correlation between the two components is zero. The sum of squares of the first component (400) is known as the first latent root and the sum of squares of the second component (100) is known as the second latent root. The f_1 is the first latent vector of W and f_2 is the second latent vector. Each PC is expressed in two different ways. The first component is expressed in terms of tests f_i and in terms of persons c_i.

(a) Determination of Matrix of Weights

The determination of the weighting matrix F starts with the computation of a matrix of sum of squares and sum of products (or correlation matrix) relating to k variables. Presuming the W matrix to be positive manifold, the first step is to obtain the sum of elements in each column. The vector of column sums is referred to as U_{a1} and when it is normalized we call it as V_{a1}.

$$V_{a1} = \frac{U_{a1}}{\sqrt{\Sigma U a_1^{\,2}}}$$

The elements in V_{a1} are accumulatively multiplied by the first row of W to obtain the first element in a new vector U_{a2}. To obtain the second element of U_{a2}, the elements of V_{a1} are accumulatively multiplied by the second row of W. The same process would be repeated for each of W and the result would be a new variable U_{a2}. Then V_{a1} and V_{a2} are compared. If they are nearly identical, then convergence is said to have occurred. Suppose that convergence

occurs when we work out V_{a8} in which case V_{a7} will be taken as the characteristic vectors. This is converted into loadings on the first PC when we multiply it by the square root of the number we obtain for normalizing U_{a8}. To obtain the second PC, the first matrix of factor cross product (Q_1) has to obtained. Then Q_1 is extracted element by element from W to obtain the first residual matrix W_1 from which the second PC are to be obtained in the same way that we obtained the first PC. The loadings of the variables, which have been reflected, have to be given negative sign. For extracting the third and subsequent PC, the same procedure outlined above has to be repeated.

The weighted matrix F is determined for the above cited example as shown below.

208	144	
144	292	
352	436	$\sqrt{314000} = 560.36$
0.6282	0.7781	
242.712	317.667	$\sqrt{159820.8} = 399.79$
0.6071	0.7946	
240.699	319.446	$\sqrt{159981.76} = 399.98 \equiv 400$
0.6	0.8	

$$f_1 = \begin{bmatrix} 0.6 \\ 0.8 \end{bmatrix} \quad \text{Hence, } F_1 = \begin{bmatrix} 0.6 \times 20 \\ 0.8 \times 20 \end{bmatrix} = \begin{bmatrix} 12 \\ 16 \end{bmatrix}$$

$$Q_1 = \begin{array}{c} \\ 12 \\ 16 \end{array} \begin{array}{c} \begin{array}{cc} 12 & 16 \end{array} \\ \begin{bmatrix} 144 & 192 \\ 192 & 256 \end{bmatrix} \end{array}$$

$$W_1 = \begin{matrix} & W & \\ \end{matrix} \quad \begin{matrix} & Q1 & \\ \end{matrix}$$

$$\begin{bmatrix} 208 & 144 \\ 144 & 292 \end{bmatrix} - \begin{bmatrix} 144 & 192 \\ 192 & 256 \end{bmatrix} = \begin{bmatrix} 64 & -48 \\ -48 & 36 \end{bmatrix}$$

After rotating the second factor,

$$W_1 = \begin{bmatrix} 64 & 48 \\ 48 & 36 \end{bmatrix}$$

112	84	$\sqrt{19600} = 140$
0.8	0.6	
80	60	$\sqrt{10000} = 100$
0.8	0.6	

$$f_2 = \begin{bmatrix} 0.8 \\ -0.6 \end{bmatrix} \quad F_2 = \begin{bmatrix} 0.8 \times 10 \\ -0.6 \times 10 \end{bmatrix} = \begin{bmatrix} 8 \\ -6 \end{bmatrix}$$

$$\text{Thus } F = \begin{bmatrix} 0.6 & 0.8 \\ 0.8 & -0.6 \end{bmatrix}$$

(b) Interpretation of the Results

The matrix F might be represented in any of the following forms. To facilitate comparisons, each form is bordered by the sum of squares of its rows and columns. Let us consider the following form denoted by F_A.

	f_1	f_2	SS
t_1	12	8	208
t_2	16	-6	292
SS	400	100	500

In this form, the elements of each latent vector have been expressed in such a way that the *sum of squares of a vector is equal to its associated root.*

Provided that all the latent vector have been included in the matrix, the sum of squares of a row is equal to the sum of squares of the test in that row.

Let us consider the following form denoted by F_B.

	f_1	f_2	SS
t_1	$\frac{12}{\sqrt{208}}$	$\frac{8}{\sqrt{208}}$	1.00
t_2	$\frac{16}{\sqrt{292}}$	$\frac{-6}{\sqrt{292}}$	1.00
SS	1.57	0.43	2.00

This form is obtained by dividing each element of F_A *by the square root of the tests sum of squares.* By squaring the appropriate element of F_B, we obtain the proportion of the total variance of a test , which is accounted for by a particular component. Thus, (144/208) = 0.69 or 69% of the variance of the first test is accounted for by the first component and the remaining 31% is accounted for by the second component. In our example, the first factor has all the loadings positive and such a factor is usually called as 'the general factor'. It is taken to represent whatever it is that all of the variables have in common. The second factor has one negative sign. Such a factor is called a `bipolar factor' and it is taken to represent a single dimension with two poles-one pole for positive loadings and the other pole for negative loadings. The row at the bottom of the table gives us further information about the usefulness of the two factors.

Let us consider the following form denoted by F_C.

	f_1	f_2	SS
t_1	$\frac{12}{\sqrt{400}}$	$\frac{8}{\sqrt{100}}$	1.00
t_2	$\frac{16}{\sqrt{400}}$	$\frac{-6}{\sqrt{100}}$	1.00
SS	1.00	1.00	2.00

This form is obtained by dividing each element of F *by the square root of the component sum of squares.* Thus the sum of squares of every column is unity. The matrix contains the normalized latent vector of W. By squaring the appropriate elements of the matrix, we obtain the proportion of a components variance which is accounted for by a particular test. The proportion of the first component of F which is attributable to the first test is 144/400 = 0.36 or 36%. The contribution of a test to a component can be immediately grasped by squaring the appropriate element of the matrix. This matrix is an orthogonal matrix. In an *orthogonal matrix*, the inner product of any row and any other row, or the inner product of any column and any other column is zero. Again, the product of any row (or column) with itself is unity.

CHAPTER 21

ANALYSIS OF THREE-DIMENSIONAL TABLES

All tables which simultaneously present data of *three qualitative variables* are known as three-dimensional contingency tables. For example, the following table shows data concerning classroom behavior of 97 school children. Three variables are involved in the table. The first variable is a teachers rating of classroom behavior classified into non-deviant and deviant. The second variable is the home risk index based on several items such as over-crowding thought to be related to deviant behavior. The third variable is an index of the adversity of school conditions based on items such as pupil turnover.

	Adversity of school conditions(k)				
	Low		High		Total
Home risk Index (j):	Not at risk	At risk	Not at risk	At risk	
Class room behavior (i):					
Non-deviant	16 (n_{111})	7 (n_{121})	20 (n_{112})	37 (n_{122})	80 ($n_{1..}$)
Deviant	1 (n_{211})	1 (n_{221})	4 (n_{212})	11 (n_{222})	17 ($n_{2..}$)
Total	17 ($n_{.11}$)	8 ($n_{.21}$)	24 ($n_{.12}$)	48 ($n_{.22}$)	97 ($n_{...}$)
Total	25 ($n_{..1}$)		72 ($n_{..2}$)		

The three-dimensional r×c×l contingency table has r rows, c columns and l layer categories. The observed frequency in the ijk^{th} cell of the table is denoted by n_{ijk} (i=1,...r; j=1,...c; k =1,...l). By summing the n_{ijk} over different subscripts, various marginal totals may be obtained. For example, summing over all values of both i and j will yield the total for the k^{th} layer category. These totals will be known as *single variable marginal totals* and we have, for example,

$$n_{i..} = \sum_{j=1}^{c} \sum_{k=1}^{r} n_{ijk}$$

Summing the n_{ijk} over any single subscript gives what we shall call the *two variable marginal totals*. For example,

$$n_{ij.} = \sum_{k=1}^{l} n_{ijk}$$

We shall analyze all *two-dimensional tables* in order to make ourselves easy to understand the results of the analysis of three-dimensional Tables. The following Table presents the data classified by classroom behavior and home risk index.

Classroom behavior	Home risk index		Total
	Not at risk	At risk	
Non-deviant	36 ($n_{11.}$)	44 ($n_{12.}$)	80 ($n_{1..}$)
Deviant	5 ($n_{21.}$)	12 ($n_{22.}$)	17 ($n_{2..}$)
Total	41 ($n_{.1.}$)	56 ($n_{.2.}$)	97 ($n...$)

$\chi^2 = 1.396$, df = 1, P > 0.05

The following table presents the data classified by classroom behavior and adversity of school conditions.

Classroom behavior	Adversity of school conditions		Total
	Low	High	
Non-deviant	23 ($n_{1.1}$)	57 ($n_{1.2}$)	80 ($n_{1..}$)
Deviant	2 ($n_{2.1}$)	15 ($n_{2.2}$)	17 ($n_{2..}$)
Total	25 ($n_{..1}$)	72 ($n_{..2}$)	97 ($n...$)

$\chi^2 = 1.320$, df = 1, $P > 0.05$

The following table presents the data classified by home risk index and adversity of school conditions.

Home risk index	Adversity of school conditions		Total
	Low	High	
Not at risk	17 ($n_{.11}$)	24 ($n_{.12}$)	41 ($n_{.1.}$)
At risk	8 ($n_{.21}$)	48 ($n_{.22}$)	56 ($n_{.2.}$)
Total	25 ($n_{..1}$)	72 ($n_{..2}$)	97 ($n...$)

$\chi^2 = 9.140$, df = 1, $P < 0.01$

Researchers with data in the form of three-dimensional contingency table should not satisfy by attempting its analysis by examining all combinations of two-dimensional tables. It may lead to misleading

conclusions because of partial and conditional independence. In case of three-dimensional tables, more than one hypothesis may be involved.

(a) Mutual Independence of Variables

The hypothesis of mutual independence of variables in a three-dimensional contingency table may be formulated as,

$$H_0 : P_{ijk} = P_{i..}\, P_{.j.}\, P_{..k}$$

Where P_{ijk} represents the probability of an observation occurring in the ijk^{th} cell, and $P_{i..}$, $P_{.j.}$ and $P_{..k}$ are the marginal probabilities of the row, column and layer variables respectively. When Ho is true, the expected values may be obtained as,

$$e_{ijk} = n\, p_{i..}\, p_{.j.}\, p_{..k}$$

Where $p_{i..}$, $p_{.j.}$ and $p_{..k}$ are estimates of the probabilities $P_{i..}$, $P_{.j.}$ and $P_{..k}$ respectively. Thus,

$$e_{ijk} = n \frac{n_{i..}}{n} \frac{n_{.j.}}{n} \frac{n_{..k}}{n}$$

$$= \frac{n_{i..}\, n_{.j.}\, n_{..k}}{n^2}$$

We compute,

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \sum_{k=1}^{l} \frac{(n_{ijk} - e_{ijk})^2}{e_{ijk}}$$

$$\text{with } df = rcl - r - c - l + 2$$

We have computed the expected values as shown in the following example.

$$e_{111} = \frac{n_{1..}\ n_{.1.}\ n_{..1}}{n^2}$$

$$= \frac{80 \times 41 \times 25}{97 \times 97} = 8.72$$

The full set of expected values are given below in the following table.

Home risk Index (j):	Adversity of school conditions (k)				Total
	Low		High		
	Not at risk	At risk	Not at risk	At risk	
Class room behavior (i):					
Non-deviant	8.72	11.90	25.10	34.28	80
Deviant	1.85	2.53	5.33	7.28	17
Total	10.57	14.43	30.43	41.56	97

$\chi^2 = 12.905$, df = 4, $P < 0.05$

(b) Partial Independence of Variables

The rejection of hypothesis of mutual independence of variables are as shown in the preceding section should not assume that there are significant associations between all variables. It might be the case that an association exists between two of the variables while the third is completely independent. In this case, hypothesis of *partial independence* would be of interest. Again, situations arise when two of the variables are independent in each level of the third, but each may be associated with this third variable. In other words, the first two variables are *conditionally independent* given the level of the third. This hypothesis may be formulated in terms of probabilities. We shall illustrate the hypothesis of *partial independence* that home risk index is associated with adversity of school conditions while classroom behaviour is

independent. That is,.

$$H_o = P_{ijk} = P_{i..}\ P_{.jk}$$

This hypothesis states that the probabilities of an observation occurring in the ijk^{th} cell (P_{ijk}) is given by the product of the probability of it falling in the i^{th} category of the row variable ($P_{i..}$) and the probability of its being in the jk^{th} cell of the column and layer classification ($P_{.jk}$). If the hypothesis is true, then it implies that the row classification is independent of both the column and the layer classification. That is, it implies the truth of the following composite hypotheses,

$$P_{ij.} = P_{i..}\ P_{.j.} \quad \text{and} \quad P_{i.k} = P_{i..}\ P_{..k}$$

The expected values are given by,

$$e_{ijk} = n\ p_{i..}\ p_{.jk}$$

Where $p_{i..}$ and $p_{.jk}$ are the estimates of the probabilities $P_{i..}$ and $P_{.jk}$ respectively. They are obtained from the relevant marginal totals as follows,

$$p_{i..} = \frac{n_{i..}}{n} \quad \text{and} \quad p_{.jk} = \frac{n_{.jk}}{n}$$

In this case, the two variables marginal totals, namely $n_{.jk}$ found by summing the observed frequencies over the first variable are needed. Using these probabilities, we obtain:

$$e_{ijk} = n\ \frac{n_{i..}}{n}\ \frac{n_{.jk}}{n}$$

$$= \frac{n_{i..}\ n_{.jk}}{n}$$

The chi-square statistics can be calculated with degrees of freedom given by,

$$df = rcl - r - cl + 1$$

In our data, this hypothesis states that the classroom behavior is *independent* of the home risk condition. The expected values are calculated. For example,

$$e_{111} = \frac{n_{1..}\ n_{.11}}{n} = \frac{80 \times 17}{97} = 14.02$$

The full set of expected values are as shown below:

	Adversity of school conditions				Total
	Low		High		
Home risk Index:	Not at risk	At risk	Not at risk	At risk	
Class room behavior :					
Non-deviant	14.02	6.60	19.80	39.59	80
Deviant	2.98	1.40	4.20	8.41	17
Total	17	8	24	48	97

$\chi^2 = 2.713, \quad df = 3, \quad P > 0.05.$

Hence we accept our hypothesis that classroom behavior is independent of the other two variables. We can see that, $e_{.jk} = n_{.jk}$. Hypothesis of partial independence fix the single variable and one set of two-variable marginal totals of the expected values to be equal to the corresponding totals of the observed value. Since we have already shown that the three variables are not mutually independent, the result leads us to conclude that there is a significant association between the school and the home conditions. We can see this by collapsing over original table into two-dimensional contingency Table as shown already. As might be predefined, fewer children from school in the low adversity category have home conditions that put them at risk for deviant behavior than would be expected if the two variables were independent.

CHAPTER 22

COMPUTERS AND ELECTRONIC DATA PROCESSING

Man knew only counting when he first began to operate with numbers. He used various devices to aid him in his counting. Probably, his fingers were the first of such devices. Later he used stones, shells, and knots. These devices were found to be inadequate and man needed a mechanical device. One such device was the counting board called as ABACUS. Pascal invented the first mechanical calculating machine in 1642. Henry Hollerith used punched cards in his mechanical data processing machines. The first computer that was based on electronic circuits was known as electronic numerical integrator and computer (ENIAC) devised by Eckert and Mauchly. These *electronic devices* enable the storage and analysis of large quantity of data at very great speed. Basic to all computer machine types is the *microprocessor*, comprising thousands of transistors on a silicon chip. The development in the production of microprocessing components and other advances in computer technology have brought about a tremendous increase in the power of computers and reduction in their size and cost. The computers may be *classified* according to the purpose for which they are devised such as analog versus digital, specific purpose versus general purpose, stored programs versus externally stored

programs,etc. The computers may also be classified as microcomputers, minicomputers and mainframes, according to power,size, and cost.

a) Computer Hardware

A computer consists of five functional units, viz., the input unit, output unit, storage unit, arithmetic-logical unit, and the central processing unit (CPU). The CPU directs the overall functions of the other units of the computer and controls the data flow between them during the process of solving a problem. It consists primarily of control circuitry which is through out the computer directing the flow of data, executing instruction operations, looking for hardware or software errors, and making sure every thing is performed at the proper times. Thus the computer consists of several physical units called as hardware. The input devices are the typewriter type of *keyboard* for entry of data into the computers. The output devices are the monitor or visual display unit (VDU), and printing/plotting units. The magnetic storage devices are the *floppy diskettes*, hard disks and cassette tapes. The floppies come in various sizes (5 ¼″, 3 ½″), densities (high density, low density) and capacities. The capacity of the computer is usually expressed in terms of 'bytes' of information it can store and handle at a time (memory).

b) Data Entry

The data processing begin after data collection has started in order to check for inconsistency, errors or incompleteness. The data is usually in the form of *unit record system* where individual records are maintained for each unit which can be arranged in any manner. Some times, the data is in the form of *registers* with appropriate columns designed for entering them in coded manner. The researcher, in conjunction with the statistician and the programmer, will prepare a list of coding

instructions which specify precisely how the data are to be converted to computer-readable form. The computers carry out not only the required analysis but also help in data entry, data editor, data management including follow-up actions, etc. The data available in these media can be read into a computer at great speeds and analyzed quickly. Once the accuracy of the fed data is ensured, the computer can be relied upon to provide error-free analysis of these data, provided the *computer program* which analyze these data is correctly written.

c) Computer Software, Statistical Packages

A special kind of programs called an operating system (OS) controls the flow of information between different components of the computer. The most popular ones are the Microsoft Disk Operating System (MS-DOS) and Unix. One of its most important jobs is to enable the hardware to understand the instructions of the software. In addition, the operating system manages files and the disks, and lists the directory of files. The use of a computer requires knowledge of the computer language such as FORTRAN (formula translators). Using these languages known as software, a program (set of instructions) is prepared for the computer to execute any given job of analysis.

The ready-made set of instructions called packages have been developed which can be used to summarize statistical data and to carry out all the common statistical analysis on these data. Special programs can be written in a suitable computer languages when pre-written package programs are not available. Some of the important statistical packages are SPSS(statistical package for social sciences), SAS(statistical analysis system), Systat, Epi-info, Statgraph and BMDP(biomedical computer program).The SPSS is the commonly used statistical packages for Biostatistics applications. The programs share procedures for manipulation and displaying data. The result is a system for analyzing

data with different methods once it is entered into the computer. The comprehensive user's manual that describe the programs is available. Most statistical programs expect to read data in a particular format. They expect to read a data set as a sequence of cards, with each card corresponding to a subject, and the cards columns corresponding to variables. The computer reads and remembers each card and then moves on to the next card until all rows have been read. Each statistical package goes about setting up its program some what differently. The data file specifies the number of variables and the format of the data. Hence the statistical packages differ considerably. Typically, this information includes the missing values and information about which variables are grouping variables etc. Finally we will need to tell the computer the name of the program. For example, the resumed word CORRM tells the computer to run the program that producers correlation coefficient matrix between each pair of variables. In each of the packages, there is considerable flexibility, and there are multiple options.

d) Versatility of Applications

Besides the applications of *statistical methods*, the computers are widely used in various aspects of mental health care research. The computer based *mental health information system* becomes a necessity. The computer based data is needed in order to formulate policy, monitor implementation and finally to evaluate it. Quality data in large quantity must be precisely calibrated to arrive in decision making. Computerized *medical records* are very efficient, quick to locate, accurate, space saving and long lasting. Computerized sheets are complete, systematic, consistent, legible, quick, less cost-wise and give a lot of information. The psychiatric medical record *linkage* provides a continuous record of individual patients from birth to death. In *hospital establishments*, computers are used to increase the efficiency and improve services to

the patients, prevent loss and drug reference system. Some times they are used in non-clinical work such as scheduling of appointments and keeping academic references. The computers undertake the evaluation on the basis of the patient information with which it is supplied and printing out its *diagnosis* or more likely, a list of possible diagnoses.

Appendix I

Protocol Writing

I. Introduction :
- (a) Formulation of problem
- (b) Review of literature
- (c) Need for the study

II. Aims & Objectives :
- (a) Primary
- (b) Secondary

III. Plan of Study :
- (a) Population to be studied-criteria for inclusion\exclusion
- (b) Investigations to be made
- (c) Standardization of terms
- (d) Census/sampling
- (e) Experimental settings
- (f) Proforma - pilot study
- (g) Personnel - their duties
- (h) Quality of data collected - cross checking etc.

IV. Plan of Analysis :
- (a) Manual/Electronic Data Processing
- (b) Statistical methods for:
  - (i) Classification, presentation & summarization of data
  - (ii) Estimation of parameters & tests of significance
  - (iii) Formation of distinct groups & identification of individuals
  - (iv) Presentation of complex multivariate data
- (c) Report writing

V. Budget

Appendix II

A Proforma to obtain Basic Data of Psychiatric Patient

1. Name of the Patient :.....
2. Date of registration : ☐☐☐☐☐☐
3. Registration number of the patient : ☐☐☐☐☐☐
4. Age (completed years) : ☐☐
5. Gender : 1. Male 2. Female ☐
6. Marital Status : 1. Unmarried 4. Divorced ☐
   2. Married 5. Separated
   3. Widow/er 6. Remarried
7. Religion : 1. Hindu 3. Christian ☐
   2. Muslim 4. Others
8. Place of Residence :
   01 - : Districts of the State ☐☐
   50 - : Neighbouring States
   60 : All other States
   61 : All other Countries
9. Domicile : 1. Urban 3. Rural ☐
   2. Semi - Urban
10. Educational : 1. Illiterate 6. Diploma / Certificate ☐
    2. Literate / Primary 7. Graduate
    3. Middle 8. Post Graduate / Professional
    4. Secondary
    5. Higher secondary PUC 9. Age : below 7 years
11. Occupation of the patient (codes given separately) : ☐☐
12. Income per month : Rs. ☐☐☐☐☐
13. Duration of illness : ☐☐☐☐☐☐
    First two columns for years
    Middle two columns for months
    Last two columns for days
14. Durations of stay in days (in-patients) : ☐☐☐
15. Follow-up attendance (for out-patients only) : ☐☐
16. Main diagnosis (ICD-10) : F ☐☐☐
17. Subsidiary diagnosis (if any) : F ☐☐☐
18. Result of treatment (% improvement) : ☐☐☐

Appendix III

THE ICD-10 CODES FOR MENTAL AND BEHAVIOURAL DISORDERS

Organic, Including Symptomatic, Mental Disorders (F00 – F09)

F00 Dementia in Alzheimer's disease

.0 . . . with early onset

.1 . . . with late onset

.2 . . . atypical or mixed type

- -

F01 Vascular dementia

.0 vascular dementia of acute onset

.1 multi-infarct dementia

.2 subcortical vascular dementia

.3 mixed cortical and subcortical vascular dementia

- -

F02 Dementia in other diseases classified elsewhere. Dementia in:

.0 . . . Pick's disease

.1 . . . Creutzfeldt-Jacob disease

.2 . . . Huntington's disease

.3 . . . Parkinson's disease

.4 . . . HIV disease

- -

F03 Unspecified dementia

F04. Organic amnesic syndrome, not induced by psychoactive substances

F05. Delirium, not induced by psychoactive substances

.0 delirium not superimposed on dementia, so described

.1 delirium superimposed on dementia

F06. **Other mental disorders due to brain damage and dysfunction, and to physical disease**

.0 organic hallucinosis

.1 organic catatonic disorder

.2 organic delusional disorder

.3 organic mood (affective) disorders

.4 organic anxiety disorder

.5 organic dissociative disorder

.6 organic emotionally labile disorder

.7 mild cognitive disorder

- -

F07 **Personality and behavioural disorders due to brain disease, damage and dysfunction**

.0 organic personality disorder

.1 postencephalitic syndrome

.2 postconcussional syndrome

- -

F09. Unspecified . . .

Mental and Behavioural Disorders due to Psychoactive Substance use (F10 – F19)

Mental and behavioural disorders due to use of:

F10. Alcohol

F11. Opioids

F12. Cannabinoids

F13. Sedatives or hypnotics

F14. Cocaine

F15. Other stimulants, including caffeine

F16. Hallucinogens

F17. Tobacco

F18. Volatile solvents

F19. Multiple drugs and other psychoactive substances

.0 acute intoxication

.1 harmful use

.2 dependence syndrome

.3 withdrawal state

.4 withdrawal state with delirium

.5 psychotic disorder

.6 amnesic syndrome

.7 residual and late-onset psychotic disorder

- -

Schizophrenia, Schizotypal and Delusional Disorders (F20 – F29)

F20. Schizophrenia

.0 paranoid . . .

.1 hebephrenic . . .

.2 catatonic . . .

.3 undifferentiated . . .

.4 post-schizophrenic depression

.5 residual. . .

.6 simple. . .

- -

F21. Schizotypal disorder

F22. Persistent delusional disorders

.0 delusional disorder

- -

F23. Acute and transient psychotic disorders

.0 acute polymorphic psychotic disorder

.1 acute polymorphic psychotic disorder with symptoms of schizophrenia

.2 acute schizophrenia-like psychotic disorder
.3 other acute predominantly delusional psychotic disorders
- -

F24. Induced delusional disorder

F25. Schizoaffective disorders
.0 . . . manic type
.1 . . . depressive type
.2 . . . mixed type
- -

F28. Other

F29. Unspecified

Mood (Affective) Disorders (F30 – F39)

F30. Manic episode
.0 hypomania
.1 mania
.2 mania with psychotic symptoms
- -

F31. Bipolar affective disorder. Current episode,
.0 . . . hypomania
.1 . . . mania
.2 . . . manic with psychotic symptoms
.3 . . . mild or moderate depression
.4 . . . severe depression
.5 . . . severe depression with psychotic symptoms
.6 . . . mixed
.7 . . . in remission
- -

F32. Depressive episode
.0 mild . . .
.1 moderate . . .
.2 severe . . .

.3 severe with psychotic symptoms

- -

F33. Recurrent depressive disorder, current episode,

.0 . . . mild

.1 . . . moderate

.2 . . . severe

.3 . . . severe with psychotic symptoms

.4 . . . in remission

- -

F34. Persistent mood (affective) disorders

.0 cyclothymia

.1 dysthymia

- -

F38. Other mood (affective) disorders

.0 other single mood (affective) disorders

.1 other recurrent mood (affective) disorders

- -

F39. Unspecified . . .

Neurotic, Stress-related and Somatoform Disorders (F40 -F48)

F40. Phobic anxiety disorders

.0 agoraphobia

.1 social phobias

.2 specific phobias

- -

F41. Other anxiety disorders

.0 panic disorder

.1 generalized anxiety disorder

.2 mixed anxiety and depressive disorder

.3 other mixed anxiety disorder

- -

F42. Obsessive-compulsive disorder

.0 predominantly obsessional thoughts or ruminations

.1 predominantly compulsive acts

.2 mixed obsessional thoughts and acts

- -

F43. **Reaction to severe stress, and adjustment disorders**

.0 acute stress reaction

.1 post-traumatic stress disorder

.2 adjustment disorders

- -

F44. **Dissociative (conversion) disorders**

.0 dissociative amnesia

.1 dissociative fugue

.2 dissociative stupor

.3 trance and possession disorders

.4 dissociative motor disorders

.5 dissociative convulsions

.6 dissociative anaesthesia and sensory loss

.7 mixed

- -

F45. **Somatoform disorders**

.0 somatization disorder

.1 undifferentiated somatoform disorder

.2 hypochondriacal disorder

.3 somatoform autonomic dysfunction

.4 persistent somatoform pain disorder

- -

F48. **Other neurotic disorders**

.0 neurasthenia

.1 depersonalization-derealization syndrome

- -

Behavioural Syndromes Associated with Physiological Disturbances and Physical Factors (F50 - F59)

F50. Eating disorders

.0 anorexia nervosa

.1 atypical anorexia nervosa
.2 bulimia nervosa
.3 atypical bulimia nervosa
.4 overeating associated with other psychological disturbances
.5 vomiting associated with other psychological disturbances
- -

F51. **Nonorganic sleep disorders**
.0 nonorganic insomnia
.1 nonorganic hypersomnia
.2 nonorganic disorder of the sleep-wake schedule
.3 sleepwalking
.4 night terrors
.5 nightmares
- -

F52. **Sexual dysfunction, not caused by organic disorder or disease**
.0 lack or loss of sexual desire
.1 sexual aversion and lack of sexual enjoyment
.2 failure of genital response
.3 orgasmic dysfunction
.4 premature ejaculation
.5 nonorganic vaginismus
.6 nonorganic dyspareunia
.7 excessive sexual drive
- -

F53. **Mental and behavioural disorders associated with the puerperium, not elsewhere classified**
.0 mild /moderate
.1 severe
- -

F54. **Psychological and behavioural factors associated with disorders or diseases classified elsewhere**

F55. **Abuse of non-dependence-producing substances**

F59. **Unspecified . . .**

Disorders of Adult Personality and Behaviour (F60 - F69)

F60. Specific personality disorders

.0 paranoid . . .
.1 schizoid . . .
.2 dissocial. . .
.3 emotionally unstable . . .
.4 histrionic . . .
.5 anankastic . . .
.6 anxious . . .
.7 dependent . . .
- -

F61. Mixed and other personality disorders

F62. Enduring personality changes, not attributable to brain damage and disease

.0 enduring personality change after catastrophic experience
.1 enduring personality change after psychiatric illness
- -

F63. Habit and impulse disorders

.0 pathological gambling
.1 pathological fire-setting
.2 pathological stealing
.3 trichotillomania
- -

F64. Gender indentity disorders

.0 transsexualism
.1 dual-role transvestism
.2 gender identity disorder of childhood
- -

F65. Disorders of sexual preference

.0 fetishism
.1 fetishistic transvestism
.2 exhibitionism
.3 voyeurism
.4 paedophilia

.5 sadomasochism
.6 multiple disorders of sexual preference
- -

F66. **Psychological and behavioural disorders associated with sexual development and orientation**
.0 sexual maturation disorder
.1 egodystonic sexual orientation
.2 sexual relationship disorder
- -

F68. **Other disorders of adult personality and behaviour**
.0 elaboration of physical symptoms for psychological reasons
.1 intentional production or feigning of symptoms or disabilities, either physical or psychological
- -

F69. **Unspecified —**

Mental Retardation (F70-F79)

F70. Mild . . .
F71. Moderate. . .
F72. Severe . . .
F73. Profound . . .
F78. Other . . .
F79. Unspecified . . .
.0 with no or minimal impairment of behaviour
.1 significant impairment of behaviour
- -

Disorders of Psychological Development (F80-F89)

F80. Specific developmental disorders of speech and language
.0 specific speech articulation disorder

.1 expressive language disorder
.2 receptive language disorder
.3 acquired aphasia with epilepsy
- -

F81. **Specific developmental disorders of scholastic skills**
.0 specific reading disorder
.1 specific spelling disorder
.2 specific disorder of arithmetical skills
.3 mixed disorder of scholastic skills
- -

F82. **Specific developmental disorder of motor function**

F83. **Mixed specific developmental disorders**

F84. **Pervasive developmental disorders**
.0 childhood autism
.1 atypical autism
.2 Rett's syndrome
.3 other childhood disintegrative disorder
.4 overactive disorder associated with MR and stereotyped movements
.5 Asperger's syndrome
- -

Behavioural and Emotional Disorders with Onset usually Occurring in Childhood and Adolescence (F90-F98)

F90. **Hyperkinetic disorders**
.0 disturbance of activity and attention
.1 hyperkinetic conduct disorder
- -

F91. **Conduct disorders**
.0 conduct disorder confined to the family context
.1 unsocialized conduct disorder

.2 socialized conduct disorder
.3 oppositional defiant disorder
- -

F92. **Mixed disorders of conduct and emotions**
.0 depressive conduct disorder
- -

F93. **Emotional disorders with onset specific to childhood**
.0 separation anxiety disorder of childhood
.1 phobic anxiety disorder of childhood
.2 social anxiety disorder of childhood
.3 sibling rivalry disorder
- -

F94. **Disorders of social functioning with onset specific to childhood and adolescence**
.0 elective mutism
.1 reactive attachment disorder of childhood
.2 disinhibited attachment disorder of childhood
- -

F95. **Tic disorders**
.0 transient tic disorder
.1 chronic motor or vocal tic disorder
.2 combined vocal and multiple motor tic disorder
- -

F98. **Other**
.0 nonorganic enuresis
.1 nonorganic encopresis
.2 feeding disorder of infancy and childhood
.3 pica of infancy and childhood
.4 stereotyped movement disorders
.5 stuttering
.6 cluttering
- -

Unspecified Mental Disorder (F99)

Appendix IV
A List of Random Numbers

1	9	6	3	8	0	1	3	7	3	3	3	0	0	5	7	1	5	4	7
0	0	1	2	0	7	7	9	0	9	3	2	1	3	3	6	3	1	5	5
6	5	7	3	0	1	1	9	5	3	8	7	4	4	5	6	4	7	4	5
3	7	6	0	9	6	4	2	1	0	8	0	8	7	2	5	5	1	0	4
1	6	1	2	6	3	9	6	2	8	4	9	3	7	6	0	4	9	0	4
5	0	9	4	5	9	9	2	9	5	0	9	1	3	6	3	4	0	9	9
5	9	7	4	5	3	0	7	8	9	5	2	2	5	5	5	6	0	2	5
8	6	8	1	4	2	6	2	1	0	4	1	2	8	7	0	5	9	8	9
0	6	1	9	6	1	5	7	2	0	3	5	7	0	4	5	0	2	4	0
3	7	1	8	4	5	7	9	7	0	5	3	1	0	1	5	3	0	6	9
3	6	5	4	1	0	3	9	1	3	3	8	6	5	7	9	2	8	0	4
8	8	1	4	1	8	9	3	8	7	5	7	9	8	1	4	7	0	5	9
6	4	5	8	5	7	2	8	5	7	8	4	2	8	3	9	1	1	6	4
7	6	2	7	1	6	4	6	8	4	6	2	5	4	3	2	5	9	2	2
4	1	4	3	5	5	9	6	0	7	6	9	5	1	3	5	0	8	7	2
1	7	6	4	9	1	0	2	2	3	6	9	9	8	2	1	1	1	6	0
9	6	2	1	6	6	6	9	5	4	2	2	1	2	1	1	9	5	9	9
0	5	1	5	1	9	3	0	8	3	3	8	9	6	5	0	8	5	0	5
6	2	2	0	3	3	9	1	4	3	5	3	6	7	5	5	0	1	0	2
9	2	7	6	9	6	0	9	4	3	2	0	4	1	0	6	2	4	9	0
5	1	5	8	4	5	6	8	5	5	7	7	0	4	7	2	1	6	9	8
0	2	9	9	0	1	0	2	3	0	7	5	2	0	6	5	9	3	2	8
9	9	7	9	6	3	9	2	3	4	4	0	2	8	0	1	3	0	2	0
3	5	3	1	5	2	1	5	8	3	1	4	0	1	1	8	3	2	2	2
8	0	9	5	4	4	3	0	5	9	2	0	6	2	3	6	5	7	2	9
5	1	9	1	9	2	4	5	6	6	3	8	4	3	6	6	2	5	5	4
4	4	0	5	7	6	5	6	3	3	1	0	6	0	6	6	6	0	7	4
0	8	1	8	2	2	9	4	2	4	3	9	2	7	1	7	3	1	8	8
7	9	2	8	6	5	9	6	3	5	3	4	0	3	3	8	3	5	9	1
8	5	7	3	9	6	7	6	8	0	9	8	0	4	8	6	0	9	5	7
8	5	6	7	3	6	7	5	7	2	0	8	6	4	3	0	6	9	0	4
7	8	5	2	0	8	0	5	4	7	4	6	2	9	8	6	3	9	1	5
4	3	5	7	2	3	0	9	1	7	4	4	6	0	6	9	9	0	3	6
2	3	1	9	7	7	2	1	5	7	3	4	5	5	6	9	0	6	6	5
7	7	1	5	0	8	8	3	6	8	8	2	1	1	6	2	1	7	1	5
5	4	0	0	9	9	6	5	6	9	7	6	2	0	4	1	5	7	9	2
2	1	9	3	6	4	6	4	6	0	7	8	8	4	0	8	8	2	9	9
5	4	2	7	8	4	7	3	6	1	7	1	8	0	9	1	1	7	8	9
1	2	4	3	0	1	1	6	0	5	6	0	3	6	9	5	5	8	1	1
7	5	1	9	3	2	9	8	3	4	2	4	7	5	4	1	6	9	1	0
8	6	1	1	5	9	6	7	3	1	0	8	5	2	7	0	8	7	4	5
6	0	4	7	4	6	9	3	0	3	8	6	1	5	6	3	5	7	0	0
2	3	6	6	5	6	4	3	4	4	6	5	6	7	2	4	3	6	4	1
9	6	9	1	9	0	5	7	1	5	9	9	3	2	2	5	2	2	9	3
9	1	7	8	8	0	4	0	4	2	0	7	8	3	2	2	2	0	9	6
6	5	1	2	6	8	2	7	1	3	0	6	6	9	1	5	7	4	8	4
6	4	1	2	1	7	6	5	2	5	3	1	6	3	4	5	2	9	7	5

Appendix V

Proportions of Areas Between Zero and Positive Values of Z in Standard Normal Distribution Curve.

Z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
.0	0.000	0.004	0.008	0.012	0.016	0.020	0.024	0.028	0.032	0.036
.1	0.040	0.044	0.048	0.052	0.056	0.060	0.064	0.068	0.071	0.075
.2	0.079	0.083	0.087	0.091	0.095	0.099	0.103	0.106	0.110	0.114
.3	0.118	0.122	0.126	0.129	0.133	0.137	0.141	0.144	0.148	0.152
.4	0.155	0.159	0.163	0.166	0.170	0.174	0.177	0.181	0.184	0.188
.5	0.192	0.195	0.199	0.202	0.205	0.209	0.212	0.216	0.219	0.222
.6	0.226	0.229	0.232	0.236	0.239	0.242	0.245	0.249	0.252	0.255
.7	0.258	0.261	0.264	0.267	0.270	0.273	0.276	0.279	0.282	0.285
.8	0.288	0.291	0.294	0.297	0.299	0.302	0.305	0.308	0.311	0.313
.9	0.316	0.319	0.321	0.324	0.326	0.329	0.332	0.334	0.337	0.339
1.0	0.341	0.344	0.346	0.349	0.351	0.353	0.355	0.358	0.360	0.362
1.1	0.364	0.367	0.369	0.371	0.373	0.375	0.377	0.379	0.381	0.383
1.2	0.385	0.387	0.389	0.391	0.393	0.394	0.396	0.398	0.400	0.402
1.3	0.403	0.405	0.407	0.408	0.410	0.412	0.413	0.415	0.416	0.418
1.4	0.419	0.421	0.422	0.424	0.425	0.427	0.428	0.429	0.431	0.432
1.5	0.433	0.435	0.436	0.437	0.438	0.439	0.441	0.442	0.443	0.444
1.6	0.445	0.446	0.447	0.448	0.450	0.451	0.452	0.453	0.454	0.455
1.7	0.455	0.456	0.457	0.458	0.459	0.460	0.461	0.462	0.463	0.463
1.8	0.464	0.465	0.466	0.466	0.467	0.468	0.469	0.469	0.470	0.471
1.9	0.471	0.472	0.473	0.473	0.474	0.474	0.475	0.476	0.476	0.477
2.0	0.477	0.478	0.478	0.479	0.479	0.480	0.480	0.481	0.481	0.482
2.1	0.482	0.483	0.483	0.483	0.484	0.484	0.485	0.485	0.485	0.486
2.2	0.486	0.486	0.487	0.487	0.488	0.488	0.488	0.488	0.489	0.489
2.3	0.489	0.490	0.490	0.490	0.490	0.491	0.491	0.491	0.491	0.492
2.4	0.492	0.492	0.492	0.493	0.493	0.493	0.493	0.493	0.493	0.494
2.5	0.494	0.494	0.494	0.494	0.495	0.495	0.495	0.495	0.495	0.495
2.6	0.495	0.496	0.496	0.496	0.496	0.496	0.496	0.496	0.496	0.496
2.7	0.496	0.497	0.497	0.497	0.497	0.497	0.497	0.497	0.497	0.497
2.8	0.497	0.498	0.498	0.498	0.498	0.498	0.498	0.498	0.498	0.498
2.9	0.498	0.498	0.498	0.498	0.498	0.498	0.498	0.499	0.499	0.499
3.0	0.499	0.499	0.499	0.499	0.499	0.499	0.499	0.499	0.499	0.499

Percentage Points of the χ^2 Distribution

df	Probability 0.05	0.01	df	Probability 0.05	0.01
1	3.84	6.63	16	26.30	32.00
2	5.99	9.21	17	27.59	33.41
3	7.82	11.34	18	28.87	34.81
4	9.49	13.28	19	30.14	36.19
5	11.07	15.09	20	31.41	37.57
6	12.59	16.81	21	32.67	38.93
7	14.07	18.48	22	33.92	40.29
8	15.51	20.09	23	35.17	41.64
9	16.92	21.67	24	36.42	42.98
10	18.31	23.21	25	37.65	44.31
11	19.68	24.72	26	38.89	45.64
12	21.03	26.22	27	40.11	46.96
13	22.36	27.69	28	41.34	48.28
14	23.69	29.14	29	42.56	49.59
15	25.00	30.58	30	43.77	50.89

Percentage Points of t-distribution for Two-tailed Tests

df	Probability 0.05	0.01	df	Probability 0.05	0.01
1	12.71	63.66	16	2.12	2.92
2	4.30	9.93	17	2.11	2.90
3	3.18	5.84	18	2.10	2.88
4	2.78	4.60	19	2.09	2.86
5	2.57	4.03	20	2.09	2.85
6	2.45	3.71	21	2.08	2.83
7	2.37	3.50	22	2.07	2.82
8	2.31	3.36	23	2.07	2.81
9	2.26	3.25	24	2.06	2.80
10	2.23	3.17	25	2.06	2.79
11	2.20	3.11	30	2.04	2.75
12	2.18	3.06	40	2.02	2.70
13	2.16	3.01	60	2.00	2.66
14	2.15	2.98	100	1.98	2.64
15	2.13	2.95	120	1.98	2.62

Percentage Points of F-distribution at 5% Level of Significance

df_n / df_d	1	2	3	4	5	6	8	12
1	161.	200.	216.	225.	230.	234.	239.	244.
2	18.51	19.00	19.16	19.25	19.30	19.33	19.37	19.41
3	10.13	9.55	9.28	9.12	9.01	8.94	8.85	8.74
4	7.71	6.94	6.59	6.39	6.26	6.16	6.04	5.91
6	5.99	5.14	4.76	4.53	4.39	4.28	4.15	4.00
8	5.32	4.46	4.07	3.84	3.69	3.58	3.44	3.28
10	4.96	4.10	3.71	3.48	3.33	3.22	3.07	2.91
12	4.75	3.89	3.49	3.26	3.11	3.00	2.85	2.69
14	4.60	3.74	3.34	3.11	2.96	2.85	2.70	2.53
16	4.49	3.63	3.24	3.01	2.85	2.74	2.59	2.42
18	4.41	3.55	3.16	2.93	2.77	2.66	2.51	2.34
20	4.35	3.49	3.10	2.87	2.71	2.60	2.45	2.28
30	4.17	3.32	2.92	2.69	2.53	2.42	2.27	2.09
40	4.08	3.23	2.84	2.61	2.45	2.34	2.18	2.00
60	4.00	3.15	2.76	2.53	2.37	2.25	2.10	1.92

Percentage Points of the F-distribution at 1% Level of Significance

df_n / df_d	1	2	3	4	5	6	8	12
1	4052.	5000.	5403.	5625.	5764.	5859.	5982.	6106.
2	98.50	99.00	99.17	99.25	99.30	99.33	99.36	99.42
3	34.12	30.82	29.46	28.71	28.24	27.91	27.50	27.05
4	21.20	18.00	16.69	16.00	15.52	15.21	14.80	14.37
6	13.75	10.90	9.78	9.15	8.75	8.47	8.10	7.72
8	11.26	8.65	7.59	7.01	6.63	6.37	6.03	5.67
10	10.00	7.56	6.55	5.99	5.64	5.39	5.06	4.71
12	9.33	6.93	5.95	5.41	5.06	4.82	4.50	4.16
14	8.86	6.51	5.56	5.04	4.69	4.46	4.14	3.80
16	8.53	6.23	5.29	4.77	4.44	4.20	3.89	3.55
18	8.29	6.01	5.09	4.58	4.25	4.01	3.71	3.37
20	8.10	5.85	4.94	4.43	4.10	3.87	3.56	3.23
30	7.56	5.39	4.51	4.02	3.70	3.47	3.17	2.84
40	7.31	5.18	4.31	3.83	3.51	3.29	2.99	2.66
60	7.08	4.98	4.13	3.65	3.34	3.12	2.82	2.50

df_n - degrees of freedom of the numerator

df_d - degrees of freedom of the denominator

Bibliography

Anderberg M R. (1975). Cluster analysis for applications. New York: Academic press.

Armitage P. (1989). Statistical methods in medical research. London: Blackwell Scientific Publications.

Bland M. (1995). An introduction to medical statistics. New York: Oxford University Press.

Cambell M J, Machin D. (1990). Medical statistics: a commonsense approach. New York: John Wiley.

Everitt B S. (1977). The analysis of contingency Tables. London: Chapman & Hall.

Fisher L D, Van Belle G (1993). Biostatistics: A methodology for the health sciences. New York: John Wiley &Sons.

Fleiss J L. (1981). Statistical methods for rates and proportions. New York: John Wiley & Sons.

Guilford J P. (1956). Fundamental statistics in psychology and education. New York:McGraw-Hill.

Hamburg M, Lubov A. (1974). Basic statistics: a modern approach. New York: Harcourt Brace Jovanovich.

Hartigan J A. (1975). Clustering algorithms. New York: John Wiley & Sons.

Hedges L V, Olkin I. (1985). Statistical methods for meta-analysis. New York: Academic Press.

Hope K. (1968). Methods of multivariate analysis. London: University of London Press.

Indian Council of Medical Research & Department of Science and Technology. (1987). Collaborative study on severe mental morbidity. New Delhi: ICMR & DST.

Kerlinger F N. (1973). Foundation of behavioural sciences. New York: Halt Reinhart & Winston.

Knappa R G. (1978). Basic statistics for nurses. New York: John Wiley & Sons.

Kothari C R.(1985) Research Methodology: Methods and Techniques. New Delhi: Wiley Eastern.

Leach C. (1979). Introduction of statistics: a non-parametric approach for the social sciences. New York: John Wiley.

Lilienfeld A M, Lilienfeld D E. (1980). Foundations of epidemiology. New York: Oxford University Press.

Lin T Y, Standley C C.(1962).The scope of epidemiology in psychiatry. Geneva:WHO Public Health Paper(16).

Lorr M. (1983). Cluster analysis for social scientists. San Francisco: Jossey-Bass.

Moser C A, Kaltron G. (1989). Survey methods in social investigations. Hants (UK): Gower Publishing Group.

Murthy M N. (1977). Sampling theory and methods. Calcutta: Statistical Publishing Society.

Registrar General of India. (1996). Population projections of India & States 1996-2016. New Delhi: RGI.

National Human Rights Commission. (1999). Quality assurance in mental health. New Delhi : NHRC.

Schmid C F, Schmid S E.(1979). Handbook of graphic presentation. New York:John Wiley & Sons.

Sharma S, Chadda R K.(1996). Mental hospitals in India: current status and role in mental health care. New Delhi: Institute of Human Behaviour & Allied Sciences.

Siegel S.(1988). Non-parametric statistics for the behavioural sciences. New York: McGraw-Hill.

Snedecor G W, Cochran W G. (1967). Statistical methods (sixth edition). Calcutta:Oxford & IBH.

Sundar Rao P S S, Richard J. (1999). An introduction to Biostatistics: a manual for students in health sciences (third edition). New Delhi: Prentice-Hall of India.

Townsend J C.(1953). Introduction to experimental method. New York: McGraw-Hill Book Company.

World Health Organization. (1993). International statistical classification of diseases and related health problems (ICD-10): classification of mental and behavioural disorders (tenth revision). Geneva: WHO.

References

Bartko J J, Carpenter W T. (1976). On the methods and theory of reliability. The Journal of Nervous and Mental Disease : 163, 307-317.

Reddy M V, Channabasavanna S M, Kaliaperumal V G. (1988). Mental health delivery system by government mental hospitals in India: NIMHANS Journal: 6, 97-106.

Reddy M V, Kaliaperumal V G, Channabasavanna S M. (1995). Mental health delivery system in general hospitals attached to medical colleges in India. Indian Journal of Psychiatry: 37, 176-178.

Reddy M V, Kaliaperumal V G, Channabasavanna S M. (1996). Mental health delivery system by government mental hospitals in India: trends during 1977-1993.NIMHANS Journal : 14, 219-222.

Reddy M V. (1996). Distribution of mental health manpower: An international scene. Indian Society for Medical Statistics Bulletin: 11, 4-7.

Reddy M V, Kaliaperumal V G, Channabasavanna S M. (1997). Cluster formation in child psychiatry – Part I: Some methodological evaluation. NIMHANS Journal: 15, 7-17.

Reddy M V, Kaliaperumal V G, Channabasavanna S M. (1997). Cluster formation in child psychiatry – Part II: Some empirical classification. NIMHANS Journal: 15, 18-30.

Reddy M V, Kapur M, Uma H. (1997). Development of a model for differentiating conduct disorder and hyperactivity in children. NIMHANS Journal: 15, 225-232.

Reddy M V, Chandrashekar C R.(1998). Prevalence of mental and behavioural disorders in India: A meta-analysis. Indian Journal of Psychiatry: 40, 149-157.

Reddy M V. (2000). Statistics, science and mental health care research. Indian Journal of Psychological Medicine: 28, 6-9.

Reddy M V. (2000). Organization and collection of data for mental health care research. Indian Journal of Psychological Medicine: 23, 10-17.

Reddy M V. (2001). A census of long-stay patients in government mental hospitals in India. Indian Journal of Psychiatry: 43, 26-34.

Reddy M V. (2001). Concept of probability and mental health care research. Indian Journal of Psychological Medicine : 24, 22-27.